(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2000－285243
（P2000－285243A）
(43)公開日　平成12年10月13日(2000.10.13)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｔ　7/00		Ｇ０６Ｆ　15/70	４６０Ａ	
Ｈ０４Ｎ　5/91		Ｈ０４Ｎ　5/91	Ｚ	
// Ｇ１０Ｌ　15/00		Ｇ１０Ｌ　3/00	５５１Ｇ	

審査請求　未請求　請求項の数37　ＯＬ　（全 27 頁）

(21)出願番号　特願2000－23339(P2000－23339)

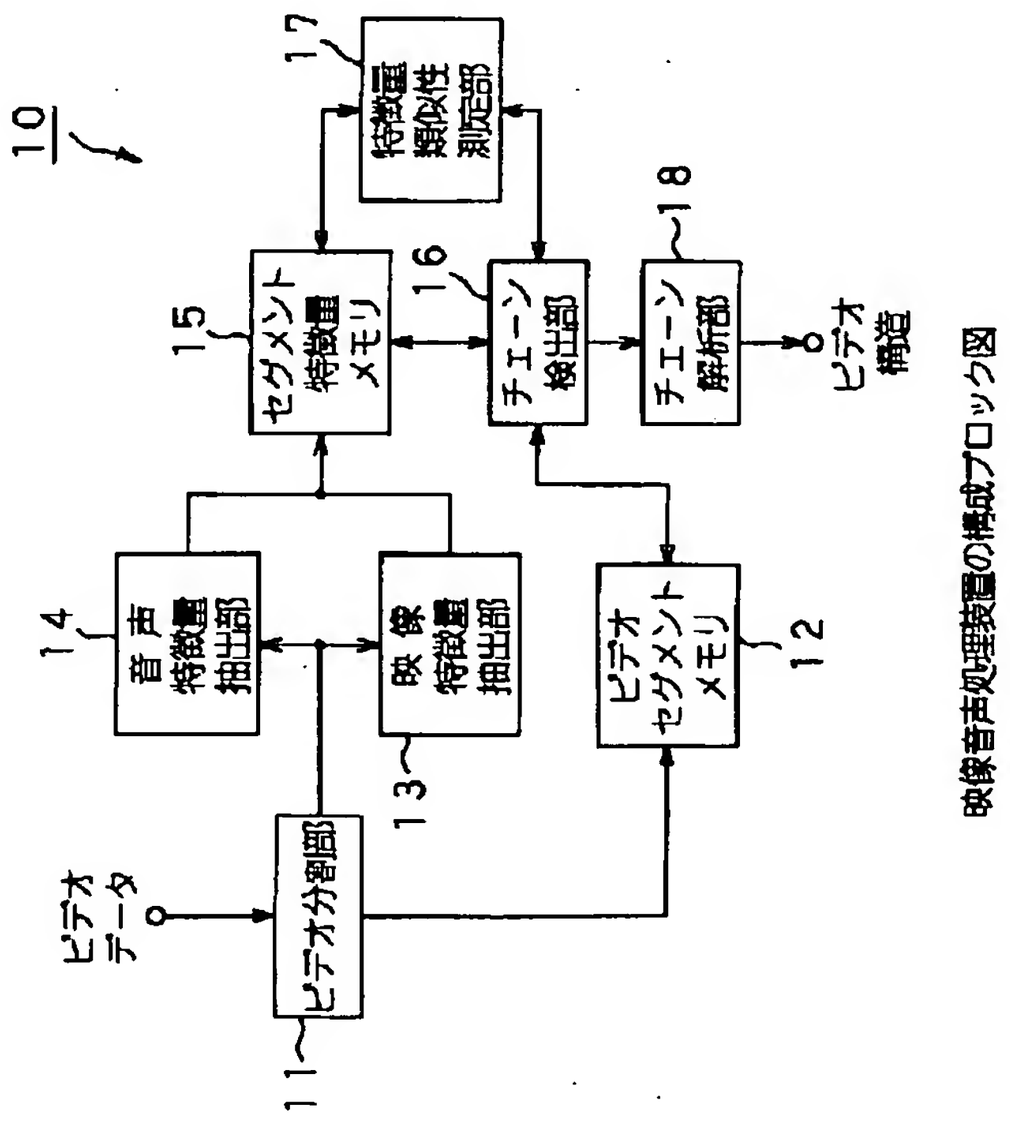
(22)出願日　平成12年１月27日(2000.1.27)

(31)優先権主張番号　特願平11－23069
(32)優先日　平成11年１月29日(1999.1.29)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000002185
ソニー株式会社
東京都品川区北品川６丁目７番35号
(72)発明者　トビー　ウォーカー
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(74)代理人　100067736
弁理士　小池　晃　（外２名）

(54)【発明の名称】　信号処理方法及び映像音声処理装置

(57)【要約】
【課題】　種々のビデオにおける高レベルのビデオ構造を抽出する。
【解決手段】　映像音声処理装置１０は、入力したビデオデータのストリームから分割された映像セグメント及び／又は音声セグメントから抽出された特徴量と、この特徴量を用いて、各特徴量毎に計算された、映像セグメント及び／又は音声セグメントの対の間の類似性を測定する測定基準とを用いて、映像セグメント及び／又は音声セグメントのうち、互いに類似する複数の映像及び／又は音声セグメントから構成される類似チェーンを検出するチェーン検出部１６と、類似チェーンを用いて解析し、ビデオの局所的ビデオ構造及び／又は大局的ビデオ構造を決定して出力するチェーン解析部１８とを備える。



映像音声処理装置の構成ブロック図

【特許請求の範囲】
【請求項１】　供給された信号の内容の意味構造を反映するパターンを検出して解析する信号処理方法であって、
上記信号を構成する連続したフレームのひと続きから形成されるセグメントから、その特徴を表す少なくとも１つ以上の特徴量を抽出する特徴量抽出工程と、
上記特徴量を用いて、上記特徴量のそれぞれ毎に、上記セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により上記セグメントの対の間の類似性を測定する類似性測定工程と、
上記特徴量と上記測定基準とを用いて、上記セグメントのうち、互いに類似する複数のセグメントから構成される類似チェーンを検出する検出工程とを備えることを特徴とする信号処理方法。
【請求項２】　上記類似チェーンを用いて解析し、上記信号の局所的構造及び／又は大局的構造を決定して出力する解析工程を備えることを特徴とする請求項１記載の信号処理方法。
【請求項３】　上記信号とは、ビデオデータにおける映像信号と音声信号との少なくとも１つであることを特徴とする請求項１記載の信号処理方法。
【請求項４】　上記類似チェーンは、当該類似チェーンが含む類似セグメントの間の関係に制約を有することを特徴とする請求項１記載の信号処理方法。
【請求項５】　上記類似チェーンは、当該類似チェーンの構造に制約を有することを特徴とする請求項１記載の信号処理方法。
【請求項６】　上記類似チェーンは、当該類似チェーンが含む全てのセグメントが互いに類似した関係にある基本類似チェーンであることを特徴とする請求項４記載の信号処理方法。
【請求項７】　上記類似チェーンは、当該類似チェーンが含む全てのセグメントにおいて、隣接するセグメントが互いに類似した関係にあるリンク類似チェーンであることを特徴とする請求項４記載の信号処理方法。
【請求項８】　上記類似チェーンは、当該類似チェーンが含む全てのセグメントにおいて、セグメントのそれぞれが、当該セグメントから所定の数だけ後方に配置されたセグメントと互いに類似した関係にある周期的チェーンであることを特徴とする請求項４記載の信号処理方法。
【請求項９】　上記類似チェーンは、当該類似チェーンが含む全てのセグメントにおいて、隣接するセグメントの各対における時間間隔が、所定の時間よりも短い局所チェーンであることを特徴とする請求項５記載の信号処理方法。
【請求項１０】　上記類似チェーンは、当該類似チェーンが含む全てのセグメントにおいて、セグメントが近似的に等時間間隔で出現する均一チェーンであることを特徴とする請求項５記載の信号処理方法。
【請求項１１】　上記検出工程は、上記特徴量と上記測定基準とを用いて、互いに類似しているセグメントを検出してまとめて候補チェーンを形成する候補チェーン検出工程と、
上記候補チェーンのそれぞれ毎に数的基準に対応する品質測定基準を算出して、上記信号の構造パターン解析における上記候補チェーンの重要性及び関連性を測定し、上記品質測定基準が所定の品質測定基準閾値を上回る候補チェーンのみを出力するフィルタリング工程とを有することを特徴とする請求項６記載の信号処理方法。
【請求項１２】　上記信号におけるセグメントのうち、セグメントが供給された時間順にしたがって当該セグメントを１つずつ逐次処理することを特徴とする請求項２記載の信号処理方法。
【請求項１３】　上記検出工程は、対象とするセグメントに関する上記特徴量と上記測定基準とを用いて、当該セグメントを含む候補チェーンを随時更新して求める候補チェーン検出工程と、
上記候補チェーンのそれぞれ毎に数的基準に対応する品質測定基準を算出して、上記信号の構造パターン解析における上記候補チェーンの重要性及び関連性を測定し、上記品質測定基準が所定の品質測定基準閾値を上回る候補チェーンのみを出力するフィルタリング工程とを有することを特徴とする請求項１２記載の信号処理方法。
【請求項１４】　上記検出工程は、周期的チェーンの初期候補を求める初期周期的チェーン検出工程と、
上記周期的チェーンの初期候補の中から、時間的に交差する重複チェーンを求める重複チェーン検出工程と、
上記重複チェーンの整合を求める整合工程とを有することを特徴とする請求項８記載の信号処理方法。
【請求項１５】　上記解析工程により、上記類似チェーンを用いて、上記信号の局所的構造として、セグメントの意味に基づく部分集合であるシーンを検出して出力することを特徴とする請求項２記載の信号処理方法。
【請求項１６】　上記解析工程により、上記類似チェーンを用いて、上記信号の大局的構造として、互いに類似するセグメントが反復的に発生する構造パターンを検出して出力することを特徴とする請求項２記載の信号処理方法。
【請求項１７】　上記構造パターンとして、ニュース放送におけるニュース項目を検出して出力することを特徴とする請求項１６記載の信号処理方法。
【請求項１８】　上記構造パターンとして、プレイが反復的に発生するスポーツ放送におけるビデオ構造を検出して出力することを特徴とする請求項１６記載の信号処理方法。
【請求項１９】　上記解析工程により、上記類似チェーンを用いて、セグメントの意味に基づく部分集合であるシーンのうち、関連するシーンをまとめたトピック構造

を検出して出力することを特徴とする請求項２記載の信号処理方法。

【請求項２０】　供給されたビデオ信号の内容の意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置であって、

上記ビデオ信号を構成する連続した映像及び／又は音声フレームのひと続きから形成される映像及び／又は音声セグメントから、その特徴を表す少なくとも１つ以上の特徴量を抽出する特徴量抽出手段と、

上記特徴量を用いて、上記特徴量のそれぞれ毎に、上記映像及び／又は音声セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により上記映像及び／又は音声セグメントの対の間の類似性を測定する類似性測定手段と、

上記特徴量と上記測定基準とを用いて、上記映像及び／又は音声セグメントのうち、互いに類似する複数の映像及び／又は音声セグメントから構成される類似チェーンを検出する検出手段とを備えることを特徴とする映像音声処理装置。

【請求項２１】　上記類似チェーンを用いて解析し、上記ビデオ信号の局所的ビデオ構造及び／又は大局的ビデオ構造を決定して出力する解析手段を備えることを特徴とする請求項２０記載の映像音声処理装置。

【請求項２２】　上記類似チェーンは、当該類似チェーンが含む類似の映像及び／又は音声セグメントの間の関係に制約を有することを特徴とする請求項２０記載の映像音声処理装置。

【請求項２３】　上記類似チェーンは、当該類似チェーンの構造に制約を有することを特徴とする請求項２０記載の映像音声処理装置。

【請求項２４】　上記類似チェーンは、当該類似チェーンが含む全ての映像及び／又は音声セグメントが互いに類似した関係にある基本類似チェーンであることを特徴とする請求項２２記載の映像音声処理装置。

【請求項２５】　上記類似チェーンは、当該類似チェーンが含む全ての映像及び／又は音声セグメントにおいて、隣接する映像及び／又は音声セグメントが互いに類似した関係にあるリンク類似チェーンであることを特徴とする請求項２２記載の映像音声処理装置。

【請求項２６】　上記類似チェーンは、当該類似チェーンが含む全ての映像及び／又は音声セグメントにおいて、映像及び／又は音声セグメントのそれぞれが、当該セグメントから所定の数だけ後方に配置された映像及び／又は音声セグメントと互いに類似した関係にある周期的チェーンであることを特徴とする請求項２２記載の映像音声処理装置。

【請求項２７】　上記類似チェーンは、当該類似チェーンが含む全ての映像及び／又は音声セグメントにおいて、隣接する映像及び／又は音声セグメントの各対における時間間隔が、所定の時間よりも短い局所チェーンであることを特徴とする請求項２３記載の映像音声処理装置。

【請求項２８】　上記類似チェーンは、当該類似チェーンが含む全ての映像及び／又は音声セグメントにおいて、映像及び／又は音声セグメントが近似的に等時間間隔で出現する均一チェーンであることを特徴とする請求項２３記載の映像音声処理装置。

【請求項２９】　上記検出手段は、上記特徴量と上記測定基準とを用いて、互いに類似している映像及び／又は音声セグメントを検出してまとめて候補チェーンを形成し、上記候補チェーンのそれぞれ毎に数的基準に対応する品質測定基準を算出して、上記ビデオ信号の構造パターン解析に対する上記候補チェーンの重要性及び関連性を測定し、上記品質測定基準が所定の品質測定基準閾値を上回る候補チェーンのみを出力することを特徴とする請求項２４記載の映像音声処理装置。

【請求項３０】　上記ビデオ信号における映像及び／又は音声セグメントのうち、映像及び／又は音声セグメントが供給された時間順にしたがって当該映像及び／又は音声セグメントを１つずつ逐次処理することを特徴とする請求項２１記載の映像音声処理装置。

【請求項３１】　上記検出手段は、対象とする上記現映像及び／又は音声セグメントに関する上記特徴量と上記測定基準とを用いて、当該映像及び／又は音声セグメントを含む候補チェーンを随時更新して求め、上記候補チェーンのそれぞれ毎に数的基準に対応する品質測定基準を算出して、上記ビデオ信号の構造パターン解析における上記候補チェーンの重要性及び関連性を測定し、上記品質測定基準が所定の品質測定基準閾値を上回る候補チェーンのみを出力することを特徴とする請求項３０記載の映像音声処理装置。

【請求項３２】　上記検出手段は、周期的チェーンの初期候補を求め、上記周期的チェーンの初期候補の中から、時間的に交差する重複チェーンを求め、上記重複チェーンの整合を求めることを特徴とする請求項２６記載の映像音声処理装置。

【請求項３３】　上記解析手段は、上記類似チェーンを用いて、上記ビデオ信号の局所的ビデオ構造として、映像及び／又は音声セグメントの意味に基づく部分集合であるシーンを検出して出力することを特徴とする請求項２１記載の映像音声処理装置。

【請求項３４】　上記解析手段は、上記類似チェーンを用いて、上記ビデオ信号の大局的ビデオ構造として、互いに類似する映像及び／又は音声セグメントが反復的に発生する構造パターンを検出して出力することを特徴とする請求項２１記載の映像音声処理装置。

【請求項３５】　上記解析手段は、上記構造パターンとして、ニュース放送におけるニュース項目を検出して出力することを特徴とする請求項３４記載の映像音声処理装置。

【請求項３６】　上記解析手段は、上記構造パターンとして、プレイが反復的に発生するスポーツ放送におけるビデオ構造を検出して出力することを特徴とする請求項３４記載の映像音声処理装置。
【請求項３７】　上記解析手段は、上記類似チェーンを用いて、映像及び／又は音声セグメントの意味に基づく部分集合であるシーンのうち、関連するシーンをまとめたトピック構造を検出して出力することを特徴とする請求項２１記載の映像音声処理装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、信号の基礎となる意味構造を反映するパターンを検出して解析する信号処理方法及びビデオ信号の基礎となる意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置に関する。
【０００２】
【従来の技術】例えばビデオデータに録画されたテレビ番組といった大量の異なる映像データにより構成される映像アプリケーションの中から、興味のある部分等の所望の部分を探して再生したい場合がある。
【０００３】このように、所望の映像内容を抽出するための一般的な技術としては、アプリケーションの主要場面を描いた一連の映像を並べて作成されたパネルであるストーリーボードがある。このストーリーボードは、ビデオデータをいわゆるショットに分解し、各ショットにおいて代表される映像を表示したものである。このような映像抽出技術は、そのほとんどが、例えば“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. ofVisual Communication and Image Representation 7:28-4, 1996”に記載されているように、ビデオデータからショットを自動的に検出して抽出するものである。
【０００４】
【発明が解決しようとする課題】ところで、例えば代表的な３０分のテレビ番組中には、数百ものショットが含まれている。そのため、上述した従来の映像抽出技術においては、ユーザは、抽出された膨大な数のショットを並べたストーリーボードを調べる必要があり、このようなストーリーボードを理解する際、ユーザに大きな負担を強いる必要があった。また、従来の映像抽出技術においては、例えば話し手の変化に応じて交互に２者を撮影した会話場面におけるショットは、冗長のものが多いという問題があった。このように、ショットは、ビデオ構造を抽出する対象としては階層が低すぎて無駄な情報量が多く、このようなショットを抽出する従来の映像抽出技術は、ユーザにとって利便のよいものとはいえなかった。
【０００５】また、他の映像抽出技術としては、例えば“A. Merlino, D. Morey and M. Maybury, Broadcast news navigation using story segmentation, Proc. of ACMMultimedia 97, 1997”や特開平１０－１３６２９７号公報に記載されているように、ニュースやフットボールゲームといった特定の内容ジャンルに関する非常に特殊な知識を用いるものがある。しかしながら、この従来の映像抽出技術は、目的のジャンルに関しては良好な結果を得ることができるものの他のジャンルには全く役に立たず、さらにジャンルに限定される結果、容易に一般化することができないという問題があった。
【０００６】さらに、他の映像抽出技術としては、例えばU.S. Patent #5,708,767号公報に記載されているように、いわゆるストーリーユニットを抽出するものがある。しかしながら、この従来の映像抽出技術は、完全に自動化されたものではなく、どのショットが同じ内容を示すものであるかを決定するために、ユーザの介入が必要であった。また、この従来の映像抽出技術は、処理に要する計算が複雑であるとともに、適用対象として映像情報のみに限定されるといった問題もあった。
【０００７】さらにまた、他の映像抽出技術としては、例えば特開平９－２１４８７９号公報に記載されているように、ショット検出と無音部分検出とを組み合わせることによりショットを識別するものがある。しかしながら、この従来の映像抽出技術は、無音部分がショット境界に対応した場合のみに限定されたものであった。
【０００８】また、他の映像抽出技術としては、例えば“H. Aoki, S. Shimotsuji and O.Hori, A shot classification method to select effective key-frames for video browsing, IPSJ Human Interface SIG Notes, 7:43-50, 1996”や特開平９－９３５８８号公報に記載されているように、ストーリーボードにおける表示の冗長を低減するために、反復された類似ショットを検出するものがある。しかしながら、この従来の映像抽出技術は、映像情報のみに適用できるものであり、音声情報に適用できるものではなかった。
【０００９】さらに、これらのような映像抽出技術は、いわゆる局所的ビデオ構造や、特殊な知識に基づく大局的ビデオ構造しか検出することができなかった。
【００１０】本発明は、このような実情に鑑みてなされたものであり、上述した従来の映像抽出技術の問題を解決し、種々のビデオデータにおける高いレベルのビデオ構造を抽出する信号処理方法及び映像音声処理装置を提供することを目的とするものである。
【００１１】
【課題を解決するための手段】上述した目的を達成する本発明にかかる信号処理方法は、供給された信号の内容の意味構造を反映するパターンを検出して解析する信号処理方法であって、信号を構成する連続したフレームのひと続きから形成されるセグメントから、その特徴を表す少なくとも１つ以上の特徴量を抽出する特徴量抽出工程と、特徴量を用いて、特徴量のそれぞれ毎に、セグメ

ントの対の間の類似性を測定する測定基準を算出して、この測定基準によりセグメントの対の間の類似性を測定する類似性測定工程と、特徴量と測定基準とを用いて、セグメントのうち、互いに類似する複数のセグメントから構成される類似チェーンを検出する検出工程とを備えることを特徴としている。

【００１２】このような本発明にかかる信号処理方法は、信号において類似したセグメントの基本的な構造パターンを検出する。

【００１３】また、上述した目的を達成する本発明にかかる映像音声処理装置は、供給されたビデオ信号の内容の意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置であって、ビデオ信号を構成する連続した映像及び／又は音声フレームのひと続きから形成される映像及び／又は音声セグメントから、その特徴を表す少なくとも１つ以上の特徴量を抽出する特徴量抽出手段と、特徴量を用いて、特徴量のそれぞれ毎に、映像及び／又は音声セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により映像及び／又は音声セグメントの対の間の類似性を測定する類似性測定手段と、特徴量と測定基準とを用いて、映像及び／又は音声セグメントのうち、互いに類似する複数の映像及び／又は音声セグメントから構成される類似チェーンを検出する検出手段とを備えることを特徴としている。

【００１４】このような本発明にかかる映像音声処理装置は、ビデオ信号において類似した映像及び／又は音声セグメントの基本的な構造パターンを決定して出力する。

【００１５】

【発明の実施の形態】以下、本発明を適用した具体的な実施の形態について図面を参照しながら詳細に説明する。

【００１６】本発明を適用した実施の形態は、録画されたビデオデータから所望の内容を自動的に探し出して抽出する映像音声処理装置である。特に、この映像音声処理装置は、ビデオデータの基礎となる意味構造を反映する映像及び／又は音声の構造パターンを検出及び解析するものであり、この解析を行うために、類似チェーン（以下、必要に応じてチェーンと略記する。）という概念を導入したものである。この映像音声処理装置の具体的な説明を行う前に、ここではまず本発明において対象とするビデオデータに関する説明を行う。

【００１７】本発明において対象とするビデオデータについては、図１に示すようにモデル化し、フレーム、セグメント、類似チェーンという構造を有するものとする。すなわち、ビデオデータは、最下位層において、一連のフレームにより構成される。また、ビデオデータは、フレームの１つ上の階層として、連続するフレームのひと続きから形成されるセグメントにより構成される。さらに、ビデオデータは、互いに特定の種類の類似パターンを有する一連のセグメントを類似チェーンとして構成する。

【００１８】このビデオデータは、映像及び音声の両方の情報を含む。すなわち、このビデオデータにおいてフレームには、単一の静止画像である映像フレームと、一般に数十～数百ミリセカンド／長といった短時間においてサンプルされた音声情報を表す音声フレームとが含まれる。

【００１９】また、セグメントは、単一のカメラにより連続的に撮影された映像フレームのひと続きから構成され、一般にはショットと呼ばれる。そして、セグメントには、映像セグメントと音声セグメントとが含まれ、ビデオ構造における基本単位となる。これらのセグメントの中で、特に音声セグメントについては、多くの定義が可能であり、例として次に示すようなものが考えられる。まず、音声セグメントは、一般によく知られている方法により検出されたビデオデータ中の無音期間により境界を定められて形成される場合がある。また、音声セグメントは、“D.Kimber and L. Wilcox, Acoustic Segmentation for Audio Browsers, Xerox Parc Technical Report”に記載されているように、例えば、音声、音楽、ノイズ、無音等のように少数のカテゴリに分類された音声フレームのひと続きから形成される場合もある。さらに、音声セグメントは、“S. Pfeiffer, S. Fischer and E. Wolfgang, Automatic Audio Content Analysis, Proceeding of ACM Multimedia 96, Nov. 1996, pp21-30”に記載されているように、２枚の連続する音声フレーム間の或る特徴における大きな変化を音声カット点として検出し、この音声カット点に基づいて決定される場合もある。

【００２０】このようなビデオデータにおいて類似チェーンとは、互いに類似し、時間的に順序付けられた複数のセグメントであって、その構造パターンは、当該チェーンに含まれる類似セグメント間の関係及びチェーンの構造として満たすべき制約条件によって、幾つかの種類に分類される。形式的には、類似チェーンとは、当該類似チェーンが含むセグメントをSi_1，・・・，Si_kで表したとき、全てのセグメントに関してｊ＝１，・・・，ｋ－１：$i_j < i_{j+1}$が成り立つ一連のセグメントである。ここで、インデックスi_jは、そのセグメントの元のビデオデータ内におけるセグメント番号を表し、ｉへの添え字ｊは、そのセグメントが当該類似チェーン内において、時間軸上でｊ番目に位置していることを表す。なお、類似チェーンには、時間的に不連続なセグメントが含まれるため、チェーンの要素間に時間的ギャップが存在することもある。換言すれば、セグメントS_{ij}，S_{ij+1}は、元のビデオデータ内において、必ずしも連続しているとは限らない。

【００２１】類似チェーンを用いることによって、ビデ

オデータにおいて、後述する局所的ビデオ構造と大局的ビデオ構造との両方に関する有力な手がかりを得ることができる。一般にビデオデータには、視聴者がその概要を知覚的に把握できる手掛かりが存在する。この手掛かりとして最も単純且つ重要なものは、類似する映像セグメント又は音声セグメントの構造パターンであり、この構造パターンこそ類似チェーンにより獲得すべき情報である。

【００２２】このような類似チェーンとしては、後に詳述するように、基本類似チェーン、リンク類似チェーン、局所チェーン、周期的チェーンがあり、これらは、ビデオデータ解析において最も重要で基本的なものである。

【００２３】ここで、基本類似チェーンとは、当該基本類似チェーンが含む全てのセグメントが互いに類似したものである。ただし、その構造パターンにおける制約はない。このような基本類似チェーンは、一般に、セグメントをグループ化するためのグルーピングアルゴリズム又はクラスタリングアルゴリズムを用いて得ることができる。また、リンク類似チェーンとは、そのチェーン内において隣接するセグメントが互いに類似したものである。さらに、局所チェーンとは、隣接するセグメントの各対において、セグメント間の時間間隔が所定の時間よりも小さいものである。そして、周期的チェーンとは、各セグメントが、それよりもm番目後方のセグメントと類似したものである。すなわち、周期的チェーンは、m個のセグメントが近似的に繰り返されることで構成される。

【００２４】そして、このような類似チェーンは、以下に示すように、ビデオデータにおける例えばシーンといった局所的ビデオ構造や、例えばニュース項目といった大局的ビデオ構造を抽出するのに用いることができる。

【００２５】ここで、シーンとは、ビデオデータを、その意味内容に基づいて、より高いレベルで記述するために、映像セグメント（ショット）検出或いは音声セグメント検出により得られたセグメントを、例えばセグメント内の知覚的アクティビティ量といったセグメントの特徴を表す特徴量を用いて意味のあるまとまりにグループ化したものである。シーンは、主観的なものであり、ビデオデータの内容或いはジャンルに依存するが、ここでは、その特徴量が互いに類似性を示す映像セグメント又は音声セグメントの反復パターンをグループ化したものとする。

【００２６】さて、上述した局所的ビデオ構造を抽出する類似チェーンの具体例として、図２に示すように、２人の話し手が互いに会話している場面において、映像セグメントが、話し手に応じて交互に現れる場合を考える。このような反復パターンを有するビデオデータにおいて、各映像セグメントは、Ａ成分及びＢ成分の各成分毎に、２つの交差するチェーンにより構成される。そのため、一般に、このような交差する局所チェーンは、関連する映像セグメントのグループ或いはシーンを検出するのに用いることができる。

【００２７】また、上述した大局的ビデオ構造を抽出する類似チェーンの具体例として、図３に示すように、固定構造を有するニュース番組を考える。このようなビデオデータにおいては、まず、各ニュース項目毎にニュースキャスターが項目を紹介するセグメントが出現し、それに続いて、例えば現地から特派員がリポートするセグメントが出現する。このような固定構造を有するビデオデータにおいては、繰り返し出現するニュースキャスターの映像セグメントは、大局的チェーンを構成する。ここで、ニュースキャスターのセグメントは、各ニュース項目の開始部を示すため、大局的チェーンを用いることによって、ニュース項目を自動的に検出することができる。すなわち、大局的チェーンを用いることによって、同図において、トピックＡ，Ｂ，Ｃ，Ｄ，・・・といった複数のニュース項目から構成されるビデオデータの中から、各トピックを検出することができる。

【００２８】本発明を適用した実施の形態として図４に示す映像音声処理装置１０は、上述したビデオデータにおけるセグメントの特徴量を用いてセグメント間の類似性を測定し、上述した類似チェーンを自動的に検出するものであり、映像セグメント及び音声セグメントの両方に適用できるものである。そして、映像音声処理装置１０は、類似チェーンを解析することによって、ビデオデータから、局所的ビデオ構造であるシーンや、大局的ビデオ構造であるトピック等の高レベルの構造を抽出・再構成することができる。

【００２９】映像音声処理装置１０は、同図に示すように、入力したビデオデータのストリームを映像、音声又はこれらの両方のセグメントに分割するビデオ分割部１１と、ビデオデータの分割情報を記憶するビデオセグメントメモリ１２と、各映像セグメントにおける特徴量を抽出する特徴量抽出手段である映像特徴量抽出部１３と、各音声セグメントにおける特徴量を抽出する特徴量抽出手段である音声特徴量抽出部１４と、映像セグメント及び音声セグメントの特徴量を記憶するセグメント特徴量メモリ１５と、映像セグメント及び音声セグメントをチェーンにまとめる検出手段であるチェーン検出部１６と、２つのセグメント間の類似性を測定する類似性測定手段である特徴量類似性測定部１７と、種々のビデオ構造を検出する解析手段であるチェーン解析部１８とを備える。

【００３０】ビデオ分割部１１は、例えば、ＭＰＥＧ１（Moving Picture Experts Group phase 1）やＭＰＥＧ２（Moving Picture Experts Group phase 2）、或いはいわゆるＤＶ（Digital Video）のような圧縮ビデオデータフォーマットを含む種々のディジタル化されたフォーマットにおける映像データと音声データとからなるビ

デオデータのストリームを入力し、このビデオデータを映像、音声又はこれらの両方のセグメントに分割するものである。このビデオ分割部１１は、入力したビデオデータが圧縮フォーマットであった場合、この圧縮ビデオデータを完全伸張することなく直接処理することができる。ビデオ分割部１１は、入力したビデオデータを処理し、映像セグメントと音声セグメントとに分割する。また、ビデオ分割部１１は、入力したビデオデータを分割した結果である分割情報を後段のビデオセグメントメモリ１２に供給する。さらに、ビデオ分割部１１は、映像セグメントと音声セグメントとに応じて、分割情報を後段の映像特徴量抽出部１３及び音声特徴量抽出部１４に供給する。

【００３１】ビデオセグメントメモリ１２は、ビデオ分割部１１から供給されたビデオデータの分割情報を記憶する。また、ビデオセグメントメモリ１２は、後述するチェーン検出部１６からの問い合わせに応じて、分割情報をチェーン検出部１６に供給する。

【００３２】映像特徴量抽出部１３は、ビデオ分割部１１によりビデオデータを分割して得た各映像セグメント毎の特徴量を抽出する。映像特徴量抽出部１３は、圧縮映像データを完全伸張することなく直接処理することができる。映像特徴量抽出部１３は、抽出した各映像セグメントの特徴量を後段のセグメント特徴量メモリ１５に供給する。

【００３３】音声特徴量抽出部１４は、ビデオ分割部１１によりビデオデータを分割して得た各音声セグメント毎の特徴量を抽出する。音声特徴量抽出部１４は、圧縮音声データを完全伸張することなく直接処理することができる。音声特徴量抽出部１４は、抽出した各音声セグメントの特徴量を後段のセグメント特徴量メモリ１５に供給する。

【００３４】セグメント特徴量メモリ１５は、映像特徴量抽出部１３及び音声特徴量抽出部１４からそれぞれ供給された映像セグメント及び音声セグメントの特徴量を記憶する。セグメント特徴量メモリ１５は、後述する特徴量類似性測定部１７からの問い合わせに応じて、記憶している特徴量やセグメントを特徴量類似性測定部１７に供給する。

【００３５】チェーン検出部１６は、ビデオセグメントメモリ１２に保持された分割情報と、１対のセグメント間の類似性とを用いて、映像セグメント及び音声セグメントをそれぞれチェーンにまとめる。チェーン検出部１６は、グループ内の各セグメントから開始して、セグメント群の中から類似しているセグメントの反復パターンを検出し、このようなセグメントをチェーンにまとめていく。このチェーン検出部１６は、チェーンの初期候補をまとめた後、第２のフィルタリング段階を用いてチェーンの最終セットを決定する。そして、チェーン検出部１６は、検出したチェーンを後段のチェーン解析部１８に供給する。

【００３６】特徴量類似性測定部１７は、２つのセグメント間の類似性を測定する。特徴量類似性測定部１７は、或るセグメントに関する特徴量を検索するようにセグメント特徴量メモリ１５に問いかける。

【００３７】チェーン解析部１８は、チェーン検出部１６により検出されたチェーン構造を解析し、種々の局所的ビデオ構造及び大局的ビデオ構造を検出する。このチェーン解析部１８は、後述するように、その細部を特定のアプリケーションに合わせて調整することができる。

【００３８】このような映像音声処理装置１０は、類似チェーンを用いて図５に概略を示すような一連の処理を行うことによって、ビデオ構造を検出する。

【００３９】まず、映像音声処理装置１０は、同図に示すように、ステップＳ１において、ビデオ分割を行う。すなわち、映像音声処理装置１０は、ビデオ分割部１１に入力されたビデオデータを映像セグメント又は音声セグメントのいずれか、或いは可能であればその両方に分割する。映像音声処理装置１０は、適用するビデオ分割方法に特に前提要件を設けない。例えば、映像音声処理装置１０は、“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているような方法によりビデオ分割を行う。このようなビデオ分割の方法は、当該技術分野ではよく知られたものであり、映像音声処理装置１０は、いかなるビデオ分割方法も適用できるものとする。

【００４０】続いて、映像音声処理装置１０は、ステップＳ２において、特徴量の抽出を行う。すなわち、映像音声処理装置１０は、映像特徴量抽出部１３や音声特徴量抽出部１４によって、そのセグメントの特徴を表す特徴量を計算する。映像音声処理装置１０においては、例えば、各セグメントの時間長、カラーヒストグラムやテクスチャフィーチャといった映像特徴量や、周波数解析結果、レベル、ピッチといった音声特徴量や、アクティビティ測定結果等が、適用可能な特徴量として計算される。勿論、映像音声処理装置１０は、適用可能な特徴量としてこれらに限定されるものではない。

【００４１】続いて、映像音声処理装置１０は、ステップＳ３において、特徴量を用いたセグメントの類似性測定を行う。すなわち、映像音声処理装置１０は、特徴量類似性測定部１７により非類似性測定を行い、その測定基準によって、２つのセグメントがどの程度類似しているかを測定する。映像音声処理装置１０は、先のステップＳ２において抽出した特徴量を用いて、非類似性測定基準を計算する。

【００４２】続いて、映像音声処理装置１０は、ステップＳ４において、チェーンの検出を行う。すなわち、映像音声処理装置１０は、先のステップＳ３において計算

した非類似性測定基準と、先のステップＳ２において抽出した特徴量とを用いて、類似したセグメントのチェーンを検出する。

【００４３】そして、映像音声処理装置１０は、ステップＳ５において、チェーンの解析を行う。すなわち、映像音声処理装置１０は、先のステップＳ４において検出したチェーンを用いて、ビデオデータの局所的ビデオ構造及び／又は大局的ビデオ構造を決定して出力する。

【００４４】このような一連の処理を経ることによって、映像音声処理装置１０は、ビデオデータからビデオ構造を検出することができる。したがって、ユーザは、この結果を用いることによって、ビデオデータの内容の索引付けや要約を行ったり、ビデオデータ中の興味のあるポイントに迅速にアクセスしたりすることが可能となる。

【００４５】以下、同図に示した映像音声処理装置１０における処理を各工程毎により詳細に説明していく。

【００４６】まず、ステップＳ１におけるビデオ分割について説明する。映像音声処理装置１０は、ビデオ分割部１１に入力されたビデオデータを映像セグメント又は音声セグメントのいずれか、或いは可能であればその両方に分割するが、このビデオデータにおけるセグメントの境界を自動的に検出するための技術は多くのものがあり、当該映像音声処理装置１０において、このビデオ分割方法に特別な前提要件を設けないことは上述した通りである。一方、映像音声処理装置１０において、後の工程によるチェーン検出の精度は、本質的に、基礎となるビデオ分割の精度に依存する。

【００４７】つぎに、ステップＳ２における特徴量抽出について説明する。特徴量とは、セグメントの特徴を表すとともに、異なるセグメント間の類似性を測定するためのデータを供給するセグメントの属性である。映像音声処理装置１０は、映像特徴量抽出部１３や音声特徴量抽出部１４により各セグメントの特徴量を計算し、セグメントの特徴を表す。映像音声処理装置１０は、いかなる特徴量の具体的詳細にも依存するものではないが、当該映像音声処理装置１０において用いて効果的であると考えられる特徴量としては、例えば以下に示す映像特徴量、音声特徴量、映像音声共通特徴量のようなものがある。映像音声処理装置１０において適用可能となるこれらの特徴量の必要条件は、非類似性の測定が可能であることである。さらに、これらの特徴量は、映像音声処理装置１０が効率化のために特徴量抽出と上述したビデオ分割とを同時に行うことを可能とする必要がある。以下に説明する特徴量は、これらの必要条件を満たすものである。

【００４８】特徴量としては、まず映像に関するものが挙げられる。以下では、これを映像特徴量と称することにする。映像セグメントは、連続する映像フレームにより構成されるため、映像セグメントから適切な映像フレームを抽出することによって、その映像セグメントの描写内容を、抽出した映像フレームで代表して表現することが可能である。すなわち、映像セグメントの類似性は、適切に抽出された映像フレームの類似性で代替可能である。このことから、映像特徴量は、映像音声処理装置１０で用いることができる重要な特徴量の１つである。この場合の映像特徴量は、単独では静的な情報しか表せないが、映像音声処理装置１０は、後述するような方法を適用することによって、この映像特徴量に基づく映像セグメントの動的な特徴を抽出することもできる。

【００４９】映像音声処理装置１０において、映像における色は、２つの映像が類似しているかを判断する際の重要な材料となる。カラーヒストグラムを用いて映像の類似性を判断することは、例えば“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているように、よく知られている。ここで、カラーヒストグラムとは、例えばＨＳＶやＲＧＢ等の３次元色空間をｎ個の領域に分割し、映像における画素の、各領域での出現頻度の相対的割合を計算したものである。そして、得られた情報からは、ｎ次元ベクトルが与えられる。圧縮されたビデオデータに関しても、例えばU.S. Patent #5,708,767号公報に記載されているように、カラーヒストグラムを、圧縮データから直接抽出することができる。

【００５０】映像音声処理装置１０では、セグメントを構成する映像におけるもともとのＹＵＶ色空間を、色チャンネル当たり２ビットでサンプルして構成した、$2^{2\cdot3}$＝６４次元のヒストグラムベクトルを用いている。

【００５１】このようなヒストグラムは、映像の全体的な色調を表すが、これには時間情報が含まれていない。そこで、映像音声処理装置１０においては、もう１つの映像特徴量として、映像相関を計算する。映像音声処理装置１０におけるチェーン検出において、複数の類似セグメントが互いに交差した構造は、それがまとまった１つのチェーン構造であることを示す有力な指標となる。例えば会話場面において、カメラの位置は、２人の話し手の間を交互に移動するが、カメラは通常、同一の話し手を再度撮影するときには、ほぼ同じ位置に戻る。このような場合における構造を検出するためには、グレイスケールの縮小映像に基づく相関がセグメントの類似性の良好な指標となることを見出したことから、映像音声処理装置１０では、元の映像をＭ×Ｎの大きさのグレイスケール映像へ間引き縮小し、これを用いて映像相関を計算する。ここで、ＭとＮは、両方とも小さな値で十分であり、例えば８×８である。すなわち、これらの縮小グレイスケール映像は、ＭＮ次元の特徴量ベクトルとして解釈される。

【００５２】さらに上述した映像特徴量とは異なる特徴

量としては、音声に関するものが挙げられる。以下では、この特徴量を音声特徴量と称することにする。音声特徴量とは、音声セグメントの内容を表すことができる特徴量であり、映像音声処理装置１０は、この音声特徴量として、周波数解析、ピッチ、レベル等を用いることができる。これらの音声特徴量は、種々の文献により知られているものである。

【００５３】まず、映像音声処理装置１０は、フーリエ変換等の周波数解析を行うことによって、単一の音声フレームにおける周波数情報の分布を決定することができる。映像音声処理装置１０は、例えば、１つの音声セグメントにわたる周波数情報の分布を表すために、ＦＦＴ（Fast Fourier Transform；高速フーリエ変換）成分、周波数ヒストグラム、パワースペクトル、その他の特徴量を用いることができる。

【００５４】また、映像音声処理装置１０は、平均ピッチや最大ピッチ等のピッチや、平均ラウドネスや最大ラウドネス等の音声レベルもまた、音声セグメントを表す有効な音声特徴量として用いることができる。

【００５５】さらに、映像音声処理装置１０は、ケプストラム特徴量として、ケプストラム係数とその１次及び２次微分係数とを含み、ＦＦＴスペクトル又はＬＰＣ（Linear Predictive Coding；線形予測符号化）等から得られたケプストラムスペクトル係数を用いることもできる。

【００５６】さらに他の特徴量としては、映像音声共通特徴量が挙げられる。これは、映像特徴量でもなく音声特徴量でもないが、映像音声処理装置１０において、チェーン内のセグメントの特徴を表すのに有用な情報を与えるものである。映像音声処理装置１０は、この映像音声共通特徴量として、アクティビティを用いる。

【００５７】アクティビティとは、セグメントの内容がどの程度動的或いは静的であるように感じられるかを表す指標である。例えば、視覚的に動的である場合、アクティビティは、カメラが対象物に沿って迅速に移動する度合い若しくは撮影されているオブジェクトが迅速に変化する度合いを表す。

【００５８】このアクティビティは、カラーヒストグラムのような特徴量のフレーム間非類似性の平均値を測定することによって、間接的に計算される。ここで、フレームｉとフレームｊとの間で測定された特徴量Ｆに対する非類似性測定基準を$d_F(i, j)$と定義すると、映像アクティビティV_Fは、次式（１）のように定義される。

【００５９】

【数１】

$$V_F = \frac{\sum_{i=b}^{f-1} d_F(i,i+1)}{f-b} \quad \cdots (1)$$

【００６０】式（１）において、ｂとｆは、それぞれ、１セグメントにおける最初と最後のフレームのフレーム番号である。映像音声処理装置１０は、具体的には、例えば上述したヒストグラムを用いて、映像アクティビティV_Fを計算することができる。

【００６１】ところで、上述した映像特徴量を始めとする特徴量は、基本的にはセグメントの静的情報を表すものであることは上述した通りであるが、セグメントの特徴を正確に表すためには、動的情報をも考慮する必要がある。そこで、映像音声処理装置１０は、以下に示すような特徴量のサンプリング方法により動的情報を表すこととする。

【００６２】映像音声処理装置１０は、例えば図６に示すように、１セグメント内の異なる時点から１以上の静的な特徴量を抽出する。このとき、映像音声処理装置１０は、特徴量の抽出数を、そのセグメント表現における忠実度の最大化とデータ冗長度の最小化とのバランスをとることにより決定する。例えば、セグメント内の或る１画像が当該セグメントのキーフレームとして指定可能な場合には、そのキーフレームから計算されたヒストグラムが、抽出すべきサンプリング特徴量となる。

【００６３】ところで、或るサンプルが常に所定の時点、例えばセグメント内の最後の時点において選択される場合を考える。この場合、黒フレームへ変化（fade）していく任意の２つのセグメントについては、サンプルが同一の黒フレームとなるため、同一の特徴量が得られる結果になる恐れがある。すなわち、これらのセグメントの映像内容がいかなるものであれ、選択した２つのフレームは、極めて類似していると判断されてしまう。このような問題は、サンプルが良好な代表値でないために発生するものである。

【００６４】そこで、映像音声処理装置１０は、このように固定点で特徴量を抽出するのではなく、セグメント全体における統計的な代表値を抽出することとする。ここでは、一般的な特徴量のサンプリング方法を２つの場合、すなわち、（１）特徴量を実数のｎ次元ベクトルとして表すことができる場合と、（２）非類似性測定基準しか利用できない場合とについて説明する。なお、（１）には、ヒストグラムやパワースペクトル等、最もよく知られている映像特徴量及び音声特徴量が含まれる。

【００６５】（１）においては、サンプル数は、事前にｋと決められており、映像音声処理装置１０は、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されてよく知られているｋ平均値

クラスタリング法（k-means-clustering method）を用いて、セグメント全体についての特徴量をｋ個の異なるグループに自動的に分割する。そして、映像音声処理装置１０は、サンプル値として、ｋ個の各グループから、グループの重心値（centroid）又はこの重心値に近いサンプルを選択する。映像音声処理装置１０におけるこの処理の複雑度は、サンプル数に関して単に直線的に増加するにとどまる。

【００６６】一方、（２）においては、映像音声処理装置１０は、“L. Kaufman and P.J.Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されているｋ－メドイドアルゴリズム法（k-medoids algorithm method）を用いて、ｋ個のグループを形成する。そして、映像音声処理装置１０は、サンプル値として、ｋ個のグループ毎に、上述したグループのメドイド（medoid）を用いる。

【００６７】なお、映像音声処理装置１０においては、抽出された動的特徴を表す特徴量についての非類似性測定基準を構成する方法は、その基礎となる静的な特徴量の非類似性測定基準に基づくが、これについては後述する。

【００６８】このようにして、映像音声処理装置１０は、静的な特徴量を複数抽出し、これらの複数の静的な特徴量を用いることによって、動的特徴を表すことができる。

【００６９】以上のように、映像音声処理装置１０は、種々の特徴量を抽出することができる。これらの各特徴量は、一般に、単一ではセグメントの特徴を表すのに不十分であることが多い。そこで、映像音声処理装置１０は、これらの各種特徴量を組み合わせることで、互いに補完し合う特徴量の組を選択することができる。例えば、映像音声処理装置１０は、上述したカラーヒストグラムと映像相関とを組み合わせることによって、各特徴量が有する情報よりも多くの情報を得ることができる。

【００７０】つぎに、図５中ステップＳ３における特徴量を用いたセグメントの類似性測定について説明する。映像音声処理装置１０は、２つの特徴量について、それがどの程度非類似であるかを測定する実数値を計算する関数である非類似性測定基準を用いて、特徴量類似性測定部１７によりセグメントの類似性測定を行う。この非類似性測定基準は、その値が小さい場合は２つの特徴量が類似していることを示し、値が大きい場合は非類似であることを示す。ここでは、特徴量Ｆに関する２つのセグメントS_1，S_2の非類似性を計算する関数を非類似性測定基準$d_F(S_1, S_2)$と定義する。このような関数は、以下の式（２）で与えられる関係を満足する。

【００７１】

【数２】

$$\begin{aligned} &d_F(S_1,S_2)=0 && (S_1=S_2\text{のとき}) \\ &d_F(S_1,S_2)\geq 0 && (\text{全ての}S_1,S_2\text{について}) \\ &d_F(S_1,S_2)=d_F(S_2,S_1) && (\text{全ての}S_1,S_2\text{について}) \end{aligned} \quad \cdots (2)$$

【００７２】ところで、非類似性測定基準の中には、或る特定の特徴量にのみ適用可能なものがあるが、“G. Ahanger and T.D.C. Little, A survey of technologies forparsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”や“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されているように、一般には、多くの非類似性測定基準は、ｎ次元空間における点として表される特徴量についての類似性を測定するのに適用可能である。その具体例は、ユークリッド距離、内積、Ｌ１距離等である。ここで、特にＬ１距離が、ヒストグラムや映像相関等の特徴量を含む種々の特徴量に対して有効に作用することから、映像音声処理装置１０は、Ｌ１距離を導入する。ここで、２つのｎ次元ベクトルをＡ，Ｂとした場合、Ａ，Ｂ間のＬ１距離d_{L1}（Ａ，Ｂ）は、次式（３）で与えられる。

【００７３】

【数３】

$$d_{L1}(A,B)=\sum_{i=1}^{n}|A_i-B_i| \quad \cdots (3)$$

【００７４】ここで、下付文字ｉは、ｎ次元ベクトルＡ，Ｂのそれぞれのｉ番目の要素を示すものである。

【００７５】また、映像音声処理装置１０は、上述したように、動的特徴を表す特徴量として、セグメントにおける様々な時点での静的な特徴量を抽出する。そして、映像音声処理装置１０は、抽出された２つの動的特徴量の間の類似性を決定するために、その非類似性測定基準として、その基礎となる静的特徴量の間の非類似性測定基準を用いる。これらの動的特徴量の非類似性測定基準は、多くの場合、各動的特徴量から選択された最も類似した静的特徴量の対の非類似性値を用いて決定されるのが最良である。この場合、２つの抽出された動的特徴量SF_1，SF_2の間の非類似性測定基準は、次式（４）のように定義される。

【００７６】

【数４】

$$d_S(SF_1, SF_2) = \min_{F_1 \in SF_1, F_2 \in SF_2} d_F(F_1, F_2) \quad \cdots (4)$$

【００７７】ここで、上式（４）における関数ｄ$_F$（Ｆ$_1$，Ｆ$_2$）は、その基礎となる静的特徴量Ｆについての非類似性測定基準を示す。なお、場合によっては、特徴量の非類似性の最小値をとる代わりに、最大値又は平均値をとってもよい。

【００７８】ところで、映像音声処理装置１０は、セグメントの類似性を決定する上で、単一の特徴量だけでは不十分であり、同一セグメントに関する多数の特徴量からの情報を組み合わせることを必要とする場合も多い。この１つの方法として、映像音声処理装置１０は、種々の特徴量に基づく非類似性を、それぞれの特徴量の重み付き組み合わせとして計算する。すなわち、映像音声処理装置１０は、ｋ個の特徴量Ｆ$_1$，Ｆ$_2$，・・・，Ｆ$_k$が存在する場合、次式（５）に表される組み合わせた特徴量に関する非類似性測定基準ｄ$_F$（Ｓ$_1$，Ｓ$_2$）を用いる。

【００７９】

【数５】

$$d_S(S_1, S_2) = \sum_{i=1}^{k} w_i d_{Fi}(S_1, S_2) \quad \cdots (5)$$

【００８０】ここで、｛ｗ$_i$｝は、Σ$_i$ｗ$_i$＝１となる重み係数である。

【００８１】以上のように、映像音声処理装置１０は、図５中ステップＳ２において抽出された特徴量を用いて非類似性測定基準を計算し、当該セグメント間の類似性を測定することができる。

【００８２】つぎに、図５中ステップＳ４におけるチェーン検出について説明する。映像音声処理装置１０は、非類似性測定基準と抽出した特徴量とを用いて、類似セグメント間のつながりを表す類似チェーンを検出する。ここでは、まず、幾つかのタイプの類似チェーンを定義し、各タイプの類似チェーンを検出するためのアルゴリズムについて具体的に説明する。

【００８３】ところで、以下に定義される類似チェーンのタイプは、それぞれ互いに独立したものであるため、映像音声処理装置１０においては、１つのチェーンが複数のタイプに属することが可能である。ここでは、このようなチェーンを、定義したタイプ名を組み合わせて称することにする。例えば、局所均一リンクチェーンは、後述するように、局所的であって均一でありリンク類似チェーンのことを示す。

【００８４】さて、類似チェーンのタイプは、当該類似チェーンが含む類似セグメント間の関係に制約を有するものと、当該類似チェーンの構造に制約を有するものとに大別される。なお、以下の定義において、チェーンＣとは、一連のセグメントＳｉ$_1$，・・・，Ｓｉ$_m$を表すこととする。ここで、インデックスｉ$_k$は、そのセグメントの、元のビデオデータ内におけるセグメント番号を表し、またｉへの添え字ｋは、そのセグメントが当該類似チェーン内において、時間軸上でｋ番目に位置していることを表す。また、これらの一連のセグメントは、常に時間軸上において順序付けられているものとし、全てのｋ＝１，・・・，ｍ－１についてｉ$_k$＜ｉ$_{k+1}$である。さらに、｜Ｃ｜は、チェーンの長さを表し、Ｃ^{start}及びＣ^{end}は、それぞれ、ビデオデータにおけるチェーンＣの開始時刻及び終了時刻を表すものとする。より正確には、チェーンＣの開始時刻は、チェーンＣにおける最初のセグメントの開始時刻であり、チェーンＣの終了時刻は、チェーンＣにおける最後のセグメントの終了時刻である。さらにまた、或るセグメントをＡとした場合、その類似セグメントを、Ａ’，Ａ’’，Ａ’’’，・・・で表す。最後に、２つのセグメントが類似しているとは、それらの非類似性測定基準が、後述する非類似性閾値よりも小さい状態であることとし、これをｓｉｍｉｌａｒ（Ｓ$_1$，Ｓ$_2$）で表す。

【００８５】当該類似チェーンが含む類似セグメント間の関係に制約を有する類似チェーンとしては、基本類似チェーン、リンク類似チェーン、周期的チェーンがある。

【００８６】まず、基本類似チェーンであるが、これは、図７に示すように、全てのセグメントが互いに類似したチェーンＣである。なお、基本類似チェーンに構造的制約はない。この基本類似チェーンは、多くの場合、類似セグメントをグループ化するためのグルーピングアルゴリズム又はクラスタリングアルゴリズムの結果として得られるものである。

【００８７】一方、リンク類似チェーンとは、図８に示すように、隣接するセグメントが互いに類似したチェーンＣである。すなわち、リンク類似チェーンでは、全てのｋ＝１，・・・，｜Ｃ｜－１について、ｓｉｍｉｌａｒ（Ｓ$_k$，Ｓ$_{k+1}$）である。このリンク類似チェーンは、上述した類似セグメントの定義から、Ａ’，Ａ’’，Ａ’’’，・・・と記述することができる。

【００８８】さらに、周期的チェーンとは、図９に示すように、各セグメントが、その後方ｍ番目のセグメントと類似したチェーンＣ$_{cyclic}$である。すなわち、周期的チェーンでは、全てのｋ＝１，・・・，｜Ｃ$_{cyclic}$｜－１について、ｓｉｍｉｌａｒ（Ｓ$_k$，Ｓ$_{k+1}$）である。換言すれば、周期的チェーンは、ｍ個の一連のセグメントの近似的な繰り返しとして構成される。これより、周期的チェーンは、Ｓ$_1$，Ｓ$_2$，・・・，Ｓ$_m$，Ｓ$_1$’，Ｓ$_2$’，・・・，Ｓ$_m$’，Ｓ$_1$’’，Ｓ$_2$’’，・・・，Ｓ$_m$’’，∧，Ｓ$_1$’’’，Ｓ$_2$’’’，・・・，Ｓ$_m$’’’と記述することができる。

【００８９】一方、構造的制約を有する類似チェーンとしては、局所チェーン、均一チェーンがある。

【００９０】ここで、局所チェーンとは、上述したように、隣接するセグメントの各対において、セグメント間

の時間間隔が所定の時間よりも小さいチェーンＣである。すなわち、局所チェーンでは、チェーン内の２つのセグメント間において許容される時間間隔の最大値をｇａｐと表すと、全てのｋ＝１，・・・，｜Ｃ｜－１について、隣接するセグメント$Ｓｉ_{k}$，$Ｓｉ_{k+1}$に対して、$ｉ_{k+1}－ｉ_{k}≦ｇａｐ$である。

【００９１】また、チェーン内のセグメントがほぼ等しい時間間隔で現れる場合、これは重要なビデオ構造の有力な指標となりうるが、このようなチェーンＣを均一チェーンと定義する。ここで、チェーンＣの均一性ｕｎｉｆｏｒｍｉｔｙ（Ｃ）を、次式（６）に示すように、等間隔時間からの時間間隔のずれの平均値を、そのチェーンの長さで規格化したものとして定義する。

【００９２】

【数６】

$$uniformity(C)=\frac{\sum_{i=1}^{|c|-1}\left|\left(S_{i+1}^{start}-S_{i}^{start}\right)-\frac{\left(C^{end}-C^{start}\right)}{|C|}\right|}{|C|\cdot\left|C^{end}-C^{start}\right|} \quad \cdots (6)$$

【００９３】上式（６）で示されるチェーンＣの均一性ｕｎｉｆｏｒｍｉｔｙ（Ｃ）は、０から１の範囲の値をとり、その値が小さい場合、セグメントの時間間隔分布が均一な分布に近いことを示す。この均一性ｕｎｉｆｏｒｍｉｔｙ（Ｃ）の値が所定の均一性閾値よりも小さい場合、チェーンＣを均一チェーンとみなす。

【００９４】以下、映像音声処理装置１０において、このような各種チェーンのそれぞれを検出するための処理について説明する。

【００９５】映像音声処理装置１０は、上述した基本類似チェーンを検出するために、バッチクラスタリング技術或いは逐次クラスタリング技術を用いる。

【００９６】バッチクラスタリング技術とは、チェーンを一括して検出する技術である。ただし、この技術を適用するためには、チェーン検出を行う前に、全てのビデオ分割を終了しておく必要がある。一方の逐次クラスタリング技術は、チェーンを逐次的に検出していく技術であり、もしビデオ分割及び特徴量抽出のまた逐次的に行われるならば、ビデオデータを再生しつつ逐次的にビデオ解析を行うことが可能となる。さらには、もし映像音声処理装置１０に十分な計算能力があるならば、この逐次的チェーン検出を実時間、換言すれば、ビデオデータを取込又は記録すると同時にチェーンを検出していくことができる。しかしながら、逐次的なビデオ解析は、その精度に問題を生じることがある。すなわち、逐次的な方法の場合、最適チェーン構造を決定するための大局的な情報がなく、さらにはセグメントの入力順序に敏感であるため、低品質の結果を生じることがある。

【００９７】映像音声処理装置１０は、バッチクラスタリング技術を用いる場合には、図１０に示すように、２つの工程を経ることによって、基本類似チェーンを検出する。

【００９８】まず、映像音声処理装置１０は、ステップＳ１１において、候補チェーンの検出を行う。すなわち、映像音声処理装置１０は、ビデオデータにおける類似セグメントを検出し、クラスタとしてまとめる。これにより得られたセグメントのクラスタ群は、基本類似チェーンを検出する上での初期候補となる。

【００９９】映像音声処理装置１０は、類似チェーンの初期候補を求める際、任意のクラスタリング技術を用いることができるが、ここでは、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されている階層的クラスタリング方法（hierarchical clustering method）を用いることにする。このアルゴリズムは、まず、最も類似した２つのセグメントを１つの対としてまとめることにより始まり、クラスタ間の類似性測定基準を用いて、各段階で最も類似したクラスタの対を次々とまとめていく。このアルゴリズムにおいて、２つのクラスタ$Ｃ_{1}$，$Ｃ_{2}$間の非類似性測定基準$ｄ_{C}$（$Ｃ_{1}$，$Ｃ_{2}$）を、次式（７）に示すように、それぞれのクラスタに含まれる２つのセグメント間の最小非類似性として定義する。

【０１００】

【数７】

$$d_{c}\left(C_{1},C_{2}\right)=\min_{s_{1}\in c_{1},s_{2}\in c_{2}} d_{s}\left(S_{1},S_{2}\right) \quad \cdots (7)$$

【０１０１】なお、映像音声処理装置１０においては、必要に応じて、上式（７）で示される最小関数の代わりに、最大関数又は平均関数を用いてもよい。

【０１０２】ところで、この階層的クラスタリング法は、仮に何らの制約のない場合、ビデオデータに含まれる全てのセグメントを単一のグループにまとめてしまう。そこで、映像音声処理装置１０は、図１１に示すように、非類似性閾値δ_{sim}を導入し、この非類似性閾値δ_{sim}との比較によって、或るセグメントが他方のセグメントと類似であるか否かを判断する。ここで、非類似性閾値δ_{sim}とは、同図に示すように、２つのセグメントがどの程度類似している場合に同一のチェーンに属するものとみなすかを決定する閾値である。そして、映像音声処理装置１０は、全クラスタ対の非類似性がこの非類似性閾値δ_{sim}を超えない範囲において、セグメント

をクラスタにまとめていく。

【０１０３】なお、映像音声処理装置１０は、非類似性閾値δ_{sim}をユーザにより設定するようにしてもよく、自動的に決定してもよい。ただし、非類似性閾値δ_{sim}として固定値を用いる場合には、その最適値は、ビデオデータの内容に依存することとなる。例えば、変化に富んだ映像内容を有するビデオデータの場合、非類似性閾値δ_{sim}は、高い値に設定される必要がある。一方、変化が少ない映像内容を有するビデオデータの場合、非類似性閾値δ_{sim}は、低い値に設定される必要がある。ここで一般に、非類似性閾値δ_{sim}が高い場合には、検出されるクラスタ数は少なくなり、非類似性閾値δ_{sim}が低い場合には、検出されるクラスタ数は多くなるという性質がある。

【０１０４】これより、映像音声処理装置１０においては、適切な非類似性閾値δ_{sim}を決定することが、その性能を左右する上で重要となる。そのため、映像音声処理装置１０においては、非類似性閾値δ_{sim}をユーザにより設定する場合には、上述したことを考慮した上で設定する必要がある。一方、映像音声処理装置１０は、以下に示す方法により、有効な非類似性閾値δ_{sim}を自動的に決定することもできる。

【０１０５】例えば、その１つの方法として、映像音声処理装置１０は、（ｎ）（ｎ－１）／２個のセグメント対の間の非類似性の分布における平均値やメジアン（中央値）といった統計量を用いて、非類似性閾値δ_{sim}を得ることができる。いま、全てのセグメント対における非類似性の平均値とその標準偏差をそれぞれμ，σとした場合、非類似性閾値δ_{sim}は、$a\mu+b\sigma$の形式で表すことができる。ここで、ａ及びｂは定数であり、それぞれ、０．５及び０．１に設定することが良好な結果を与えることを見出している。

【０１０６】実用上においては、映像音声処理装置１０は、全てのセグメント対について、それらの間の非類似性を求める必要はなく、その平均値μ及び標準偏差σが真値に十分近い結果を与えるに足りるセグメント対を、全セグメント対集合からランダムに選択し、その非類似性を求めればよい。映像音声処理装置１０は、このようにして得られた平均値μ及び標準偏差σを用いることによって、適切な非類似性閾値δ_{sim}を自動的に得ることができる。すなわち、映像音声処理装置１０は、例えば、セグメント対の全数をｎ、任意の小さい定数をＣとした場合、Ｃｎで与えられる数のセグメント対の非類似性を抽出することによって、適切な非類似性閾値δ_{sim}を自動的に決定することができる。

【０１０７】映像音声処理装置１０は、これまでに示したようにセグメントのクラスタリングを行った後、各クラスタにて、当該各クラスタに含まれるセグメントを並べ替えることによって、基本類似チェーンの初期候補を得ることができる。

【０１０８】ところで、図１０中ステップＳ１１において検出したチェーン候補は、その多くが、実際のビデオ構造とは無関係のものである。これより、映像音声処理装置１０は、どのチェーン候補がビデオ構造の骨格をなす重要なチェーンであるか、或いは、ビデオ構造に関連するチェーンであるかを決定する必要がある。そのため、映像音声処理装置１０は、ステップＳ１２において、チェーンの品質を示す数的基準に対応する品質測定基準を用いたチェーンフィルタリングを行う。すなわち、映像音声処理装置１０は、ビデオ構造解析におけるチェーン候補の重要性及び関連性を測定し、所定の品質測定基準閾値を上回るチェーン候補のみをチェーン検出の結果として出力する。ここで、フィルタリングで使用される関連性測定関数として最も単純な例は、チェーン候補が受け入れられるか否かを示すブール関数であるが、映像音声処理装置１０は、必要に応じて、より複雑な関連性測定関数を用いてもよい。

【０１０９】ところで、映像音声処理装置１０においては、チェーン品質測定基準として、チェーン長、チェーン密度、チェーン強度等が用いられる。

【０１１０】まず、チェーン長であるが、これは、１つのチェーンが保有するセグメントの数と定義される。ここで、映像音声処理装置１０が、このチェーン長を、そのチェーン品質測定基準として用いることができるのは、一般にチェーン長が小さい場合であり、それは通常ノイズとしてみなすことが可能であることに依る。例えば、或るチェーンが単一セグメントしか有していない場合、それは何らの情報を有していない。すなわち、チェーン長に基づく品質測定基準では、その制約として、チェーンが保有すべきセグメント数の最小値が与えられることとなる。

【０１１１】次に、チェーン密度であるが、これは、或るチェーンが保有する全セグメント数と、そのチェーンが占めるビデオデータの部分領域における全セグメント数との比として定義される。これは、チェーンが限られた時間領域内に集中して存在する方が好ましい場合があることに依る。この場合、映像音声処理装置１０は、このチェーン密度を、そのチェーン品質測定基準として用いればよい。

【０１１２】最後に、チェーン強度であるが、これは、チェーン内の各セグメントが互いにどの程度類似しているかを示す指標であり、当該セグメントが互いに類似しているほど、そのチェーンは高い強度を有しているとみなす。なお、映像音声処理装置１０において、このチェーン強度を測定する方法については、以下に示すチェーン内類似性測定法や、全ての可能なセグメント対の間の非類似性の平均値をとる方法、或いは、全ての可能なセグメント対の間の非類似性の最大値をとる方法を含め、多数存在する。

【０１１３】一例として、映像音声処理装置１０が、チ

ェーン内類似性測定法によりチェーン強度を測定する場合を示す。ここで、チェーン内類似性測定法とは、チェーンを構成するセグメントの類似性を、それぞれのセグメントと、そのチェーンが含む最も代表的なセグメントとの非類似性の平均値として表す方法である。典型的なセグメントの例としては、チェーンの重心（centroid）セグメントが挙げられる。いま、チェーンＣにおける重心セグメントを$S_{centroid}$とすると、この重心セグメント$S_{centroid}$は、次式（８）で定義される。

【０１１４】

【数８】

$$S_{centroid} = \underset{S_A \in C}{argmin} \frac{1}{|C|} \sum_{S_B \in C} d_F(S_A, S_B) \quad \cdots (8)$$

【０１１５】ここで、上式（８）におけるａｒｇｍｉｎは、評価対象の式の値を最小とする入力$S_A \in C$を選択することを表す。

【０１１６】これより、チェーン強度を$d_{centroid}$とすると、このチェーン強度$d_{centroid}$は、次式（９）のように表される。

【０１１７】

【数９】

$$d_{centroid} = \frac{1}{|C|} \sum_{S \in C} d_F(S, S_{centroid}) \quad \cdots (9)$$

【０１１８】さて、映像音声処理装置１０は、上述したチェーン品質測定基準を用いて、具体的に図１２に示すような一連の処理によりチェーンフィルタリングを行う。

【０１１９】まず、映像音声処理装置１０は、ステップＳ２１において、チェーンリストC_{list}を候補チェーンで初期化するとともに、フィルタリングチェーンリスト$C_{filtered}$を空状態にする。

【０１２０】続いて、映像音声処理装置１０は、ステップＳ２２において、チェーンリストC_{list}が空状態であるか否かを判別する。

【０１２１】ここで、チェーンリストC_{list}が空状態であった場合には、映像音声処理装置１０は、対象とする候補チェーンが存在しないことから、一連の処理を終了する。

【０１２２】一方、チェーンリストC_{list}が空状態でない場合には、映像音声処理装置１０は、ステップＳ２３において、或るチェーンＣをチェーンリストC_{list}の最初の要素とし、チェーンＣをチェーンリストC_{list}から除去する。

【０１２３】続いて、映像音声処理装置１０は、ステップＳ２４において、チェーンＣに関するチェーン品質測定基準を計算する。

【０１２４】そして、映像音声処理装置１０は、ステップＳ２５において、このチェーン品質測定基準が品質測定基準閾値よりも大きいか否かを判別する。

【０１２５】ここで、チェーン品質測定基準が品質測定基準閾値よりも小さい場合には、映像音声処理装置１０は、ステップＳ２２へと処理を移行し、再び別のチェーンに関する処理を行う。

【０１２６】一方、チェーン品質測定基準が品質測定基準閾値よりも大きい場合には、映像音声処理装置１０は、ステップＳ２６において、フィルタリングチェーンリスト$C_{filtered}$にチェーンＣを追加する。

【０１２７】そして、映像音声処理装置１０は、ステップＳ２７において、チェーンリストC_{list}が空状態であるか否かを判別する。

【０１２８】ここで、チェーンリストC_{list}が空状態であった場合には、映像音声処理装置１０は、対象とする候補チェーンが存在しないことから、一連の処理を終了する。

【０１２９】一方、チェーンリストC_{list}が空状態でない場合には、映像音声処理装置１０は、ステップＳ２３へと処理を移行する。このようにして、映像音声処理装置１０は、チェーンリストC_{list}が空状態となるまで処理を繰り返す。

【０１３０】このような一連の処理によって、映像音声処理装置１０は、チェーンフィルタリングを行い、どのチェーンが、ビデオ構造の骨格をなす重要なチェーンであるか、或いは、ビデオ構造に関連するチェーンであるかを決定することができる。

【０１３１】以上のように、映像音声処理装置１０は、このようなバッチクラスタリング技術を用いて、基本類似チェーンを検出することができる。

【０１３２】ところで、映像音声処理装置１０は、バッチクラスタリング技術とは別の方法として、上述した逐次クラスタリング技術を用いて、基本類似チェーンを検出することもできる。すなわち、映像音声処理装置１０は、ビデオデータにおけるセグメントを、その入力の順にしたがって１つずつ処理して、チェーン候補リストを繰り返し更新していく。映像音声処理装置１０は、この場合にも、バッチクラスタリング技術と同様に、チェーン検出の主たる工程を２段階に分けて行う。すなわち、映像音声処理装置１０は、まず、逐次クラスタリングアルゴリズムを用いて、類似セグメントのクラスタを検出する。次に、映像音声処理装置１０は、バッチクラスタリング技術と同様のチェーン品質測定基準を用いて、検出されたクラスタをフィルタリングしていく。ここで、

映像音声処理装置１０は、逐次クラスタリング技術を用いた場合のフィルタリング処理として、チェーンのフィルタリングが早い段階で進められる点において、バッチクラスタリング技術の場合と異なる。

【０１３３】さて、逐次クラスタリング技術においては、セグメントのクラスタリングを行う際に、逐次クラスタリングアルゴリズムを用いる。ところで、一般に、ほとんどの逐次クラスタリングは、局所最適に行われる。すなわち、逐次クラスタリングアルゴリズムでは、新たなセグメントが入力される度に、そのセグメントを既存のクラスタに割り当てるか、或いは、そのセグメントのみを含む新たなクラスタを生成するかを局所的に判断している。一方、より精巧な逐次クラスタリングアルゴリズムとしては、セグメントの入力順序にともなうバイアス効果を防ぐため、新たなセグメントが入力される度に、クラスタ分割そのものを更新するものもある。このようなアルゴリズムについては、“J. Roure and L. Talavera, Robust incremental clustering with bad instance orderings:a new strategy,In Proceedings of the Sixth Iberoamerican Conference on Artifical Intelligence, IBERAMIA-98. Pages 136-147. Lisbon, Portugal. Helder Coelho ed.,LNAI vol. 1484. Springer Verlag, 1998”の記載を参照することができる。

【０１３４】映像音声処理装置１０は、逐次クラスタリングアルゴリズムの一例として、図１３に示すような処理を行う。ここでは、セグメントに分割されたビデオデータが、セグメントS_1、・・・、S_nを有しているものとする。なお、ここでは、チェーン解析の工程も含めた一連の処理について説明する。

【０１３５】まず、映像音声処理装置１０は、同図に示すように、ステップＳ３１において、チェーンリストC_{list}を空状態に初期化し、ステップＳ３２において、セグメント番号ｉを１に設定する。

【０１３６】次に、映像音声処理装置１０は、ステップＳ３３において、セグメント番号ｉが総セグメント数ｎよりも小さいか否かを判別する。

【０１３７】ここで、セグメント番号ｉが総セグメント数ｎよりも大きい場合には、映像音声処理装置１０は、対象とするセグメントが存在しないため、一連の処理を終了する。

【０１３８】一方、セグメント番号ｉが総セグメント数ｎよりも小さい場合には、映像音声処理装置１０は、ステップＳ３４において、セグメントS_1、すなわちここではセグメントS_1を取り込み、ステップＳ３５において、チェーンリストC_{list}が空状態であるか否かを判別する。

【０１３９】ここで、チェーンリストC_{list}が空状態である場合には、映像音声処理装置１０は、ステップＳ４２へと処理を移行する。

【０１４０】一方、チェーンリストC_{list}が空状態でない場合には、映像音声処理装置１０は、ステップＳ３６において、セグメントS_1に対する非類似性が最小であるチェーンC_{min}を求める。ここで、チェーンC_{min}は、次式（１０）のように定義される。

【０１４１】

【数１０】

$$C_{\min} = \underset{C \in C_{list}}{argmin}\, d_{SC}(C, S_i) \quad \cdots (10)$$

【０１４２】上式（１０）において、d_{SC}（Ｃ，Ｓ）は、チェーンＣとセグメントＳとの間の非類似性測定基準を表し、次式（１１）で与えられる。

【０１４３】

【数１１】

$$d_{SC}(C, S) = \underset{s_1 \in C}{min}\, d_F(S, S_1) \quad \cdots (11)$$

【０１４４】これは、バッチクラスタリング技術において定義した類似性測定基準である上式（７）において、その第２引数を、当該セグメントのみを含んだクラスタとしたものと等価である。以下では、チェーンC_{min}とセグメントS_iとの間の最小非類似性d_{SC}（C_{min}, S_i）を、単にd_{min}として表すこととする。

【０１４５】次に、映像音声処理装置１０は、ステップＳ３７において、バッチクラスタリング技術の場合において説明したような非類似性閾値δ_{sim}を用い、最小非類似性d_{min}が非類似性閾値δ_{sim}よりも小さいか否かを判別する。

【０１４６】ここで、最小非類似性d_{min}が非類似性閾値δ_{sim}よりも大きい場合には、映像音声処理装置１０は、ステップＳ４２の処理へと移行し、唯一の要素セグメントとして当該セグメントS_iのみを有する新たなチェーンC_{new}を生成し、ステップＳ４３において、新たなチェーンC_{new}をチェーンリストC_{list}に追加して、ステップＳ３９の処理へと移行する。

【０１４７】一方、最小非類似性d_{min}が非類似性閾値δ_{sim}よりも小さい場合には、映像音声処理装置１０は、ステップＳ３８において、チェーンC_{min}に当該セグメントS_iを追加する。すなわち、映像音声処理装置１０は、$C_{min} \leftarrow C_{min} \cup S_i$とする。

【０１４８】そして、映像音声処理装置１０は、ステップＳ３９において、チェーンをフィルタリングする。すなわち、映像音声処理装置１０は、上述したように、各要素チェーン$C \in C_{list}$について、チェーンＣの品質を測定して、品質測定基準閾値を上回る品質測定基準を有

するチェーンのみを選択し、これをチェーンリストC$_{filtered}$に追加する。

【０１４９】さらに、映像音声処理装置１０は、ステップＳ４０において、逐次的にチェーンを解析する。すなわち、映像音声処理装置１０は、その時点でのフィルタリングされたチェーンリストC$_{filtered}$を解析モジュールに通す。

【０１５０】そして、映像音声処理装置１０は、ステップＳ４１において、セグメント番号ｉに１を加算し、ステップＳ３３の処理へと移行する。

【０１５１】このようにして、映像音声処理装置１０は、セグメント番号ｉが総セグメント数ｎよりも大きくなるまで、以上の一連の処理を繰り返し、セグメント番号ｉが総セグメント数ｎよりも大きくなった際のチェーンリストC$_{list}$の各要素チェーンを、基本類似チェーンとして検出する。

【０１５２】なお、同図に示す一連の処理は、入力されたビデオデータに含まれる総セグメント数ｎが既知であることを前提としている。しかしながら、一般には、総セグメント数ｎが前もって与えられていない場合も多い。その場合、逐次クラスタリングアルゴリズムは、同図中ステップＳ３３において、セグメントの入力が引き続きあるか否かによって、処理の続行或いは終了を判別すればよい。

【０１５３】このような一連の処理によって、映像音声処理装置１０は、逐次クラスタリング技術を用いた基本類似チェーンの検出を行うことができる。

【０１５４】つぎに、上述したリンク類似チェーンを検出する処理について説明する。映像音声処理装置１０における、リンク類似チェーンの検出は、基本類似チェーン検出の特殊なケースとして考えることができる。映像音声処理装置１０は、逐次クラスタリングアルゴリズムを用いたリンク類似チェーン検出方法として、図１４に示すような処理を行う。ここでは、セグメントに分割されたビデオデータが、セグメントS$_1$、・・・、S$_n$を有しているものとする。なお、ここでは、チェーン解析の工程も含めた一連の処理を説明する。

【０１５５】映像音声処理装置１０は、同図に示すように、ステップＳ５１において、チェーンリストC$_{list}$を空状態に初期化し、ステップＳ５２において、セグメント番号ｉを１に設定する。

【０１５６】次に、映像音声処理装置１０は、ステップＳ５３において、セグメント番号ｉが総セグメント数ｎよりも小さいか否かを判別する。

【０１５７】ここで、セグメント番号ｉが総セグメント数ｎよりも大きい場合には、映像音声処理装置１０は、対象とするセグメントが存在しないため、一連の処理を終了する。

【０１５８】一方、セグメント番号ｉが総セグメント数ｎよりも小さい場合には、映像音声処理装置１０は、ステップＳ５４において、セグメントS$_1$、すなわちここではセグメントS$_1$を取り込み、ステップＳ５５において、セグメントS$_1$に対する非類似性が最小であるチェーンC$_{min}$を求める。ここで、チェーンC$_{min}$は、次式（１２）のように定義される。

【０１５９】

【数１２】

$$C_{\min} = \underset{c \in C_{list}}{argmin}\, d_{SC}(C, S_i) \quad \cdots (12)$$

【０１６０】上式（１２）において、d$_{SC}$（Ｃ，Ｓ）は、やはりチェーンＣとセグメントＳとの間の非類似性測定基準を表すが、リンク類似チェーン検出においては、この非類似性測定基準d$_{SC}$（Ｃ，Ｓ）は、次式（１３）で与えられる。

【０１６１】

【数１３】

$$d_{SC} = (C, S)\, d_F(S_{|C|}, S_i) \quad \cdots (13)$$

【０１６２】すなわち、非類似性測定基準d$_{SC}$（Ｃ，Ｓ）は、基本類似チェーンの検出の際に用いた非類似性測定基準である上式（１１）とは異なり、当該セグメントと、チェーンＣにおける最後の要素セグメントとの間の非類似性として与えられる。

【０１６３】次に、映像音声処理装置１０は、ステップＳ５６において、上述したような非類似性閾値δ$_{sim}$を用い、最小非類似性d$_{min}$が非類似性閾値δ$_{sim}$よりも小さいか否かを判別する。

【０１６４】ここで、最小非類似性d$_{min}$が非類似性閾値δ$_{sim}$よりも大きい場合には、映像音声処理装置１０は、ステップＳ６１の処理へと移行し、唯一の要素セグメントとして当該セグメントS$_i$のみを有する新たなチェーンC$_{new}$を生成し、ステップＳ６２において、新たなチェーンC$_{new}$をチェーンリストC$_{list}$に追加して、ステップＳ５８の処理へと移行する。

【０１６５】一方、最小非類似性d$_{min}$が非類似性閾値δ$_{sim}$よりも小さい場合には、映像音声処理装置１０は、ステップＳ５７において、チェーンC$_{min}$の末端に当該セグメントS$_i$を追加する。すなわち、映像音声処理装置１０は、C$_{min}$←C$_{min}$，S$_i$とする。

【０１６６】そして、映像音声処理装置１０は、ステップＳ５８において、チェーンをフィルタリングする。すなわち、映像音声処理装置１０は、上述したように、各要素チェーンC∈C$_{list}$について、チェーンＣの品質を測定して、品質測定基準閾値を上回る品質測定基準を有

するチェーンのみを選択し、これをチェーンリストC$_{filtered}$に追加する。なお、映像音声処理装置１０は、この工程を省略することもできる。

【０１６７】さらに、映像音声処理装置１０は、ステップＳ５９において、逐次的にチェーンを解析する。すなわち、映像音声処理装置１０は、その時点でのフィルタリングされたチェーンリストC$_{filtered}$を解析モジュールに通す。

【０１６８】そして、映像音声処理装置１０は、ステップＳ６０において、セグメント番号ｉに１を加算し、ステップＳ５３の処理へと移行する。

【０１６９】このようにして、映像音声処理装置１０は、セグメント番号ｉが総セグメント数ｎよりも大きくなるまで、以上の一連の処理を繰り返し、セグメント番号ｉが総セグメント数ｎよりも大きくなった際のチェーンリストC$_{list}$の各要素チェーンを、リンク類似チェーンとして検出する。

【０１７０】このような一連の処理によって、映像音声処理装置１０は、このような逐次クラスタリング技術を用いて、リンク類似チェーンを検出することができる。

【０１７１】なお、同図に示す一連の処理は、入力されたビデオデータに含まれる総セグメント数ｎが既知であることを前提としている。しかしながら、一般には、総セグメント数ｎが前もって与えられていない場合も多い。その場合、逐次クラスタリングアルゴリズムは、同図中ステップＳ５３において、セグメントの入力が引き続きあるか否かによって、処理の続行或いは終了を判別すればよい。

【０１７２】つぎに、上述した周期的チェーンを検出する処理について説明する。周期的チェーンC$_{cyclic}$は、ｋ個の異なる基本類似チェーン又はリンク類似チェーンがまとまったもの｛C$_1$，・・・，C$_k$｝とみなすことができる。以下、周期的チェーンC$_{cyclic}$内のセグメントを、S$_1$，・・・，S$_n$と記述し、またC（S$_i$）は、セグメントS$_i$の出現元のチェーン番号１、・・・、ｋを示すこととする。これより、C$_{cyclic}$が周期的チェーンであるならば、C（S$_1$），C（S$_2$），・・・，C（S$_n$）なる一連のチェーン番号の並びは、i$_1$，・・・，i$_k$，i$_1$，・・・，i$_k$，・・・，i$_1$，・・・，i$_k$という形式で記述されることとなる。ここで、その１周期分i$_1$，・・・，i$_k$は、チェーン番号１，・・・，ｋの順列、換言すれば、重複しない任意の並びである。なお、以下では、１周期内に含まれるセグメントの数が１つである周期的チェーンi$_1$，i$_1$，・・・，i$_1$を基本周期チェーンと称することとする。

【０１７３】ところで、通常、ビデオデータにおける周期的構造は、各周期が完全に一致したものではなく近似的なものであるため、映像音声処理装置１０は、図１５に示すような一連の処理によって、ビデオデータ内の近似的な周期的チェーンを探す。ここで、映像音声処理装置１０は、必要に応じて、その元となる基本周期チェーンが均一でなければならないという制約条件を追加することができる。ここでは、この制約条件のもとに行われる処理について説明する。

【０１７４】まず、映像音声処理装置１０は、同図に示すように、ステップＳ７１及びステップＳ７２において、ビデオデータに含まれる基本周期チェーンを検出し、それに基づいて初期チェーンリストを生成し、さらに初期チェーンリストに含まれる基本周期チェーンの全てが均一チェーンの制約条件を満たすように、初期チェーンリストを更新する。

【０１７５】すなわち、映像音声処理装置１０は、ステップＳ７１において、上述した基本類似チェーン又はリンク類似チェーンを検出するアルゴリズムを用いて、初期チェーンリストC$_{list}$を求める。

【０１７６】そして、映像音声処理装置１０は、ステップＳ７２において、初期チェーンリストに含まれる各チェーンCについて、その均一性を確認し、チェーンCが均一でない場合には、このチェーンCを、その時間的間隔が最大となるような複数の均一サブチェーンに分割する。続いて、映像音声処理装置１０は、得られた均一サブチェーンを、上述した基本類似チェーン又はリンク類似チェーンを検出するアルゴリズムにおいて説明したようなチェーン品質測定基準を用いてフィルタリングし、選択された均一サブチェーンを初期チェーンリストC$_{list}$に追加する。

【０１７７】次に、映像音声処理装置１０は、ステップＳ７３において、チェーンリストC$_{list}$の中から、時間的に重複して交差する１対のチェーン、すなわち、∃C$_1$，C$_2$｜［C$_1^{start}$，C$_1^{end}$］∩［C$_2^{start}$，C$_2^{end}$］なるチェーンC$_1$，C$_2$を求める。

【０１７８】そして、映像音声処理装置１０は、ステップＳ７４において、このような重複しているチェーンC$_1$，C$_2$が存在するか否かを判別する。

【０１７９】ここで、重複しているチェーンC$_1$，C$_2$が存在しない場合には、映像音声処理装置１０は、チェーンリストC$_{list}$が既に複数の周期的チェーンを含んでいるものとして、一連の処理を終了する。

【０１８０】一方、重複しているチェーンC$_1$，C$_2$が存在する場合には、映像音声処理装置１０は、ステップＳ７５乃至ステップＳ７８において、２つのチェーンC$_1$，C$_2$がまとまった１つの周期的チェーンを構成するか否かを決定するため、その２つの周期的チェーンを合わせた周期的チェーンにおいて、各周期の間の整合性を評価する。

【０１８１】すなわち、映像音声処理装置１０は、ステップＳ７５において、２つのチェーンC$_1$，C$_2$を合わせて、新たな周期的チェーンC$_M$を形成する。ここで、チェーンC$_M$におけるセグメントをS$_1$，S$_2$，・・・，S$_{|CM|}$と表すこととする。

【０１８２】続いて、映像音声処理装置１０は、ステップＳ７６において、セグメントS_1の出現元のチェーン番号Ｃ（S_1）をＣとし、チェーン番号の並びＣ（S_1），Ｃ（S_2），・・・，Ｃ（$S_{|CM|}$）においてＣの発生毎に、すなわち、セグメントS_1と同じチェーンに属するセグメントが出現する直前を境に、チェーンＣ

$$C_M^1 = S_1, \cdots, S_i,$$
$$C_M^2 = S_{i_1+1}, \cdots, S_{i_2},$$
$$\vdots$$
$$C_M^k = S_{i_{k-1}+1}, \cdots, S_{i_k},$$

【０１８４】この操作から明らかなように、上式（１４）では、全ての$C_M{}^j$について、Ｃ（S_{ij+1}）＝Ｃ（S_1）が成り立つ。

【０１８５】続いて、映像音声処理装置１０は、ステップＳ７７において、最も出現頻度の高いサブチェーンＣ$_M$をサブチェーン$C_M{}^1$，$C_M{}^2$，・・・，$C_M{}^k$に分解する。この結果、映像音声処理装置１０は、次式（１４）に示すようなサブチェーンのリストを得る。

【０１８３】

【数１４】

$$C_M^{cycle} = \underset{C_M^k}{argmax} \left| \left\{ C_M^i \middle| C_M^i = C_M^k, i \in \{1, \cdots, k\} \right\} \right| \quad \cdots (15)$$

【０１８７】そして、映像音声処理装置１０は、ステップＳ７８において、最も出現頻度の高いサブチェーン$C_M{}^{cycle}$が、元のチェーンＣ$_M$の１周期となりうるか否かを評価する。すなわち、映像音声処理装置１０は、整合係数ｍｅｓｈを、次式（１６）で示すように、ステップＳ７６にて求めた$C_M{}^{cycle}$の出現頻度のサブチェーン総数に対する比で定義し、続くステップＳ７９において、この整合係数が所定の閾値を越えるか否かを判別する。

【０１８８】

【数１６】

$$mesh = \frac{\left| \left\{ C_M^i \middle| C_M^i = C_M^{cycle}, i \in \{1, \cdots, k\} \right\} \right|}{k} \quad \cdots (16)$$

【０１８９】ここで、整合係数が閾値を越えていない場合には、映像音声処理装置１０は、ステップＳ７３の処理へと移行し、他の重複しているチェーンを求めて同様の処理を繰り返す。

【０１９０】一方、整合係数が閾値を越えている場合には、映像音声処理装置１０は、ステップＳ８０において、チェーンC_1，C_2をチェーンリストC_{list}から除去して、ステップＳ８１において、チェーンＣ$_M$をチェーンリストC_{list}に追加し、ステップＳ７３の処理へと移行する。

【０１９１】映像音声処理装置１０は、チェーンリストC_{list}に含まれる全ての周期的チェーンについて重複しているチェーンが存在しなくなるまでこのような一連の処理を繰り返すことによって、最終的な周期的チェーンを含むチェーンリストC_{list}を得ることができる。

【０１９２】以上のように、映像音声処理装置１０は、非類似性測定基準と抽出した特徴量とを用いて、類似したセグメントの各種チェーンを検出することができる。

【０１９３】つぎに、図５中ステップＳ５におけるチェーン解析について説明する。映像音声処理装置１０は、検出したチェーンを用いて、ビデオデータの局所的ビデオ構造及び／又は大局的ビデオ構造を決定して出力する。ここでは、ビデオデータに発生する基本的な構造パターンを検出するのに、チェーン解析の結果をどのように用いるのかについて具体的な例を挙げて説明する。

【０１９４】まず、ビデオデータに発生する局所的な構造パターンであるシーンについて説明する。

【０１９５】シーンは、上述したように、セグメントのレベルより上位に位置づけられた最も基本的な局所的ビデオ構造の単位であり、意味的に関連する一連のセグメントから構成される。映像音声処理装置１０は、チェーンを用いて、これらのシーンを検出することができる。映像音声処理装置１０におけるシーン検出において、チェーンが満たすべき条件とは、そのチェーンが含む全てのセグメントに関して、互いに連続したセグメント間の時間間隔が、時間閾値と称される或る定められた値を超えないことである。ここでは、この条件を満たすチェーンを局所チェーンと称する。

【０１９６】映像音声処理装置１０は、チェーンを用いてシーンを検出するために、図１６に示すような一連の

処理を行う。
【０１９７】まず、映像音声処理装置１０は、同図に示すように、ステップＳ９１乃至ステップＳ９４において、局所チェーンリストを求める。
【０１９８】すなわち、映像音声処理装置１０は、ステップＳ９１において、上述した基本類似チェーン検出アルゴリズムを用いて、１組の初期チェーンリストを求める。
【０１９９】次に、映像音声処理装置１０は、ステップＳ９２において、求めた初期チェーンリストにおける各チェーンＣについて、チェーンＣが局所チェーンでない場合には、チェーンＣを、局所チェーンの条件範囲において最長であるところの局所サブチェーン$C=C_1, \cdots, C_n$の並びに分解する。
【０２００】その後、映像音声処理装置１０は、ステップＳ９３において、チェーンリストからチェーンＣを除去する。
【０２０１】さらに、映像音声処理装置１０は、ステップＳ９４において、各サブチェーンC_iをチェーンリストに追加する。この工程が終了すると、全てのチェーンが局所的となる。
【０２０２】次に、映像音声処理装置１０は、ステップＳ９５において、チェーンリストの中から、時間的に交差する１対の重複しているチェーンC_1，C_2、すなわち、$\exists C_1, C_2 \mid [C_1^{start}, C_1^{end}] \cap [C_2^{start}, C_2^{end}]$であるところのチェーン$C_1$，$C_2$を求める。
【０２０３】続いて、映像音声処理装置１０は、ステップＳ９６において、このような重複しているチェーンC_1，C_2が存在するか否かを判別する。
【０２０４】ここで、重複しているチェーンC_1，C_2が存在しない場合には、映像音声処理装置１０は、チェーンリストに含まれた各チェーン毎に１つのシーンが存在するものとして、一連の処理を終了する。
【０２０５】一方、重複しているチェーンC_1，C_2が存在する場合には、映像音声処理装置１０は、ステップＳ９７において、重複しているチェーンC_1，C_2を合わせて、新たなチェーンC_Mを形成する。
【０２０６】さらに、映像音声処理装置１０は、ステップＳ９８において、チェーンリストから重複しているチェーンC_1，C_2を除去して、チェーンC_Mを追加し、その後再びステップＳ９５の処理へと移行して、同様の処理を繰り返す。
【０２０７】このようにした結果、重複しているチェーンがチェーンリスト内に存在しなくなったとき、最終的に得られたチェーンリストに含まれた各チェーン毎に、１シーンが存在することになる。なお、チェーンC_Jに対応するシーンS_Jの境界は、C^{start}及びC^{end}で与えられる。
【０２０８】ところで、セグメントの中には、いかなるチェーンにも割り当てられずに残るものがあるが、映像音声処理装置１０は、既定値としては、２つの検出されたシーン間に残ったこのようなセグメントをまとめて１つのシーンとする。
【０２０９】このような一連の処理によって、映像音声処理装置１０は、チェーンを用いることによって、ビデオデータにおける局所的な構造パターンであるシーンを検出することができる。
【０２１０】このような処理を先に図２に示した会話場面に適用する場合を考える。この場合、映像音声処理装置１０は、ステップＳ９１乃至ステップＳ９４において、話し手のセグメントのそれぞれについて、局所チェーンを求める。そして、映像音声処理装置１０は、ステップＳ９７において、これらのチェーンをまとめ、シーン全体を表す単一の大きいチェーンを形成することになる。
【０２１１】このように、映像音声処理装置１０は、会話場面におけるシーンを検出することができる。
【０２１２】なお、映像音声処理装置１０においては、シーンを検出した際に、シーン内の全てのセグメントがチェーンに含まれる訳ではないことには注意を要する。
【０２１３】また、映像音声処理装置１０は、上述したアルゴリズムを逐次的に行うことによって、シーンを逐次的に検出することもできる。
【０２１４】つぎに、大局的な構造パターンとして、ニュース項目を検出する場合について説明する。
【０２１５】上述したように、ニュース番組は、そのニュース項目が、例えば、まずアンカーによる導入文で始まり、現場からの１以上のリポートが続くといった周期的構造を有している。すなわち、このようなビデオ構造は、アンカーショットから次のアンカーショットの直前までを１周期とした単純な周期的構造であるとみなすことができる。
【０２１６】映像音声処理装置１０は、チェーンを用いてニュース項目を自動的に検出するために、図１７に概略を示すような一連の処理を行う。
【０２１７】まず、映像音声処理装置１０は、同図に示すように、ステップＳ１０１において、上述した周期的チェーン検出アルゴリズムを用いて、周期的チェーンの検出を行う。この工程を行うことによって、映像音声処理装置１０は、周期的チェーンのリストを得ることができる。ここで、各周期は、ニュース項目を表してもよく、表さなくてもよい。
【０２１８】次に、映像音声処理装置１０は、ステップＳ１０２において、その周期が、ビデオデータの全長の所定割合よりも短いところの周期的チェーンを全て除去する。すなわち、映像音声処理装置１０は、この工程を行うことによって、ニュース項目を表す見込みのない短い周期の周期的チェーンを排除することができる。このような周期は、例えば司会者がゲストにインタビューをする場合或いは他の短時間周期がニュース放送において

現れる場合に発生しうるものである。
【０２１９】そして、映像音声処理装置１０は、ステップＳ１０３において、ステップＳ１０２において残った全ての周期的チェーンについて、時間的に最も短い周期的チェーンを求め、この周期的チェーンが他の周期的チェーンに重なる場合には、その周期的チェーンを周期的チェーンのリストから除去する。映像音声処理装置１０は、いかなる周期的チェーンも他の周期的チェーンと重なることがなくなるまで、この処理を繰り返す。このステップＳ１０３が終了した後に残った周期的チェーンのリストは、検出したニュース項目リストを含むこととなる。すなわち、ステップ１０３にて得られた周期的チェーンのリストの各周期は、それぞれ、１つのニュース項目を表す。
【０２２０】このようにして、映像音声処理装置１０は、チェーンを用いてニュース項目を自動的に検出することができる。
【０２２１】なお、特筆すべきは、映像音声処理装置１０は、例えば、ニュース放送のメイン、スポーツ、ビジネスの各セグメントの間といったニュース放送の途中にニュースキャスターが変わった場合にも、問題なく作用することができることである。
【０２２２】つぎに、スポーツ放送におけるプレイを検出する場合について説明する。
【０２２３】多くのスポーツは、同じ一連の工程が何度も繰り返されることによりプレイが構成されるといった固定パターンを有するという特徴がある。例えば、野球の場合には、ピッチャーがボールを投げ、バッターがボールを打とうとすることによりプレイが構成される。ビデオデータにおいて、このようなプレイ構造を有する他のチームスポーツとしては、例えばフットボールやラグビーが挙げられる。
【０２２４】このプレイ構造が放送されると、ビデオデータは、プレイの各部分についてのセグメント群の繰り返しを表すこととなる。すなわち、ビデオデータは、ピッチャーを表すセグメントの後に、バッターを表すセグメントが続き、ボールが打たれた場合には、外野選手等を表すセグメントが入ることになる。そのため、野球放送に対して映像音声処理装置１０によるチェーン検出を適用した場合には、ビデオデータにおいて、ピッチャーを表すセグメントが１チェーンとして検出され、バッターを表すセグメントが別の１チェーンを占め、その他のチェーンが外野や種々の光景にあたることになる。
【０２２５】すなわち、これらのスポーツ放送においては、プレイ構造が、上述した周期的チェーン検出方法を用いて検出することができる周期的映像となる。このような他の例として、テニスが挙げられる。テニスにおいて、ビデオデータは、サーブ、ボレー、サーブ、ボレーといったような周期を構成する。この場合、各サーブを表すセグメントは、映像的に互いに類似しているため、映像音声処理装置１０は、プレイを検出するために、このようなセグメントを用いることができる。その結果、映像音声処理装置１０による構造解析においては、近似的にゲームのプレイ構造を検出することができる。
【０２２６】さらに、他のスポーツ、特に個人競技においては、プレイ構造としては、１人の競技者が或る活動を完結するまで行うことになるが、各競技者は、全て近似的に同じ活動を行っているとみなすことができる。例えば、スキージャンプ競技では、各競技者が１回ジャンプを行い、次の競技者が続いて同様のジャンプを行う。すなわち、ジャンプ競技の放送におけるビデオデータは、競技者がジャンプの準備に入り、助走路を滑り降りて、着地するというセグメントの並びからなるのが一般的である。これより、ビデオデータは、このような一連のセグメントを、各競技者毎に繰り返すことで構成される。このような放送におけるビデオデータに対してチェーン検出を適用した場合には、ジャンプの各段階毎に類似した一連のチェーンを検出することになる。したがって、各競技者毎の周期は、周期的チェーン検出方法を用いて抽出することができる。
【０２２７】映像音声処理装置１０において、チェーン解析によりスポーツ放送におけるプレイを自動的に検出する際には、適当でないチェーンを排除するために、さらなる制約を設ける必要がある場合がある。どのような制約が適切であるかは、スポーツの種類によって異なるが、例えば、映像音声処理装置１０は、検出された周期的チェーンのうち、その周期が十分長いものだけをプレイとして検出するという経験的なルールを用いることができる。
【０２２８】すなわち、映像音声処理装置１０は、チェーンを用いてスポーツ放送におけるプレイを自動的に検出するために、図１８に概略を示すような一連の処理を行う。
【０２２９】まず、映像音声処理装置１０は、同図に示すように、ステップＳ１１１において、上述した周期的チェーン検出アルゴリズムを用いて、周期的チェーンを検出する。
【０２３０】そして、映像音声処理装置１０は、ステップＳ１１２において、得られたチェーンのリストに対して品質条件を適用し、そのチェーンリストをフィルタリングして、本質的でないチェーンを除去する。品質条件としては、例えば、プログラムの大部分にわたるような周期的チェーンのみを残すといったことが挙げられる。勿論、映像音声処理装置１０は、対象とするスポーツに特有の制約条件を追加してもよい。
【０２３１】このようにして、映像音声処理装置１０は、チェーン解析によりスポーツ放送におけるプレイを自動的に検出することができる。
【０２３２】つぎに、周期検出とシーン検出とを組み合わせてトピックを検出する場合について説明する。

【０２３３】例えば、ドラマ、コメディ、バラエティといった多くのテレビ番組におけるビデオデータは、上述したシーンにより構成されている。しかし、ビデオデータは、その上位の構造として、幾つかの関連シーンの並びから構成されるトピックなる構造を有する場合がある。このトピックは、必ずしも、常にスタジオ司会者による紹介セグメントに始まるようなニュース放送におけるトピックと類似したものであるとは限らない。例えば、視覚的な例として、紹介セグメントの代わりにロゴイメージのセグメント或いは総合司会者のセグメントが用いられたり、或いは聴覚的な例として、新たなトピックが始まる度に、常に同じテーマ音楽が流されたりする場合がある。

【０２３４】或る番組におけるビデオデータが、このようなトピック構造を有しているか否かは、周期検出とシーン検出とを組み合わせることによって、判断することができる。

【０２３５】そのため、映像音声処理装置１０は、チェーンを用いた周期検出とシーン検出とを組み合わせたトピック検出を行うために、図１９に概略を示すような一連の処理を行う。

【０２３６】まず、映像音声処理装置１０は、同図に示すように、ステップＳ１２１において、基本類似チェーン検出を行い、１組の基本類似チェーンリストを識別する。

【０２３７】次に、映像音声処理装置１０は、ステップＳ１２２において、周期的チェーン検出を行い、１組の周期的チェーンのリストを識別する。

【０２３８】続いて、映像音声処理装置１０は、ステップＳ１２３において、ステップＳ１２１において求めた基本類似チェーンリストを用い、先に図１６に示したアルゴリズムを適用して、シーン構造を抽出する。映像音声処理装置１０は、この結果、シーンのリストを得ることができる。

【０２３９】そして、映像音声処理装置１０は、ステップＳ１２４において、ステップＳ１２２において求めた周期的チェーンのリストを、ステップＳ１２３において検出した各シーン要素と比較する。ここで、映像音声処理装置１０は、検出したシーンのリストに含まれるシーンよりも短い周期の周期的チェーンを全て除去する。この結果得られた残りの周期的チェーンは、各周期が幾つかのシーンを有しているが、この各周期はそれぞれ、候補トピックとして識別されることとなる。

【０２４０】このようにして、映像音声処理装置１０は、チェーンを用いた周期検出とシーン検出とを組み合わせることによって、トピック検出を行うことができる。

【０２４１】なお、映像音声処理装置１０は、ステップＳ１２４において、その他の制約や品質条件を設けることによって、トピック検出の精度を高めることもできる。

【０２４２】以上のように、映像音声処理装置１０は、検出した各種チェーンを用いて、ビデオデータの各種局所的ビデオ構造及び／又は各種大局的ビデオ構造を決定して出力することができる。

【０２４３】以上説明してきたように、本発明の実施の形態として示す映像音声処理装置１０は、互いに類似する複数の映像セグメント又は音声セグメントから構成される類似チェーンを検出することが可能である。そして、映像音声処理装置１０は、これらの類似チェーンを解析することによって、高レベルのビデオ構造を抽出することができる。特に、映像音声処理装置１０は、局所的ビデオ構造及び大局的ビデオ構造の解析を共通の枠組みで行うことができる。

【０２４４】この映像音声処理装置１０は、完全に自動的に処理を行うことができ、ユーザが事前にビデオデータの内容の構造を知る必要はない。

【０２４５】また、映像音声処理装置１０は、逐次的なチェーン検出を用いることにより、逐次的にビデオ構造を解析することも可能であり、さらに、プラットホームの計算能力が十分強力であるならば、ビデオ構造解析を実時間で行うことが可能である。これにより、映像音声処理装置１０は、事前に記録されたビデオデータの他、ライブのビデオ放送にも用いることができる。例えば、映像音声処理装置１０は、スポーツ放送におけるプレイ検出において、ライブのスポーツ放送に適用可能である。

【０２４６】さらに、映像音声処理装置１０は、ビデオ構造を検出した結果、ビデオブラウジングのための新たな高レベルアクセスの基礎を与えることができる。すなわち、映像音声処理装置１０は、セグメントではなくトピックといった高レベルのビデオ構造を用いてビデオデータの内容を映像化することにより、内容に基づいたビデオデータへのアクセスを可能とする。例えば、映像音声処理装置１０は、シーンを表示することにより、ユーザは、番組の要旨をすばやく知ることができ、興味のある部分を迅速に見つけることができる。

【０２４７】さらにまた、映像音声処理装置１０は、ニュース放送におけるトピック検出の結果を用いることにより、ユーザに、ニュース項目単位での選択や視聴を可能とする等、ニュース放送に対して、強力で新しい方法のアクセスを可能とする。

【０２４８】また、映像音声処理装置１０は、ビデオ構造検出の結果、ビデオデータの要約を自動的に作成するための基礎を与えることができる。一般に、筋の通った要約を作成するためには、ビデオデータに含まれる任意のセグメントを組み合わせるのではなく、ビデオデータを再構成可能な意味を持つ成分に分解し、それを元に適切なセグメントを組み合わせることが必要である。映像音声処理装置１０により検出されたビデオ構造は、その

ような要約を作成するための基礎的な情報を提供するものである。

【０２４９】さらに、映像音声処理装置１０では、ビデオデータを、そのジャンル別に解析することが可能である。例えば、映像音声処理装置１０は、テニスの試合のみを検出することを可能とする。

【０２５０】これより、映像音声処理装置１０は、放送局におけるビデオ編集システムに組み込まれることにより、ビデオデータを、その内容に基づいて編集することを可能とする。

【０２５１】さらにまた、映像音声処理装置１０は、一般家庭において、ホームビデオを解析したり、ホームビデオからビデオ構造を自動的に抽出するのに用いることができる。さらに、映像音声処理装置１０は、ビデオデータの内容の要約や、その内容に基づいた編集を行うのに用いることができる。

【０２５２】一方、映像音声処理装置１０は、ビデオチェーンを、人手によるビデオデータの内容の解析を補足するツールとして使用することが可能である。特に、映像音声処理装置１０は、チェーン検出の結果を映像化することにより、ビデオデータの内容のナビゲーションやビデオ構造解析が容易にすることができる。

【０２５３】また、映像音声処理装置１０は、そのアルゴリズムが非常に単純であり計算上の効率もよいため、セットトップボックスやディジタルビデオレコーダ、ホームサーバ等の家庭用電子機器にも適用することができる。

【０２５４】なお、本発明は、上述した実施の形態に限定されるものではなく、例えば、セグメント間の類似性測定のために用いる特徴量や、適用可能なビデオデータの内容等は、上述したもの以外でもよいことは勿論であり、その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。

【０２５５】

【発明の効果】以上詳細に説明したように、本発明にかかる信号処理方法は、供給された信号の内容の意味構造を反映するパターンを検出して解析する信号処理方法であって、信号を構成する連続したフレームのひと続きから形成されるセグメントから、その特徴を表す少なくとも１つ以上の特徴量を抽出する特徴量抽出工程と、特徴量を用いて、特徴量のそれぞれ毎に、セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準によりセグメントの対の間の類似性を測定する類似性測定工程と、特徴量と測定基準とを用いて、セグメントのうち、互いに類似する複数のセグメントから構成される類似チェーンを検出する検出工程とを備える。

【０２５６】したがって、本発明にかかる信号処理方法は、信号において類似したセグメントが構成する基本的な構造パターンを検出することができ、これらの構造パターンがどのように組み合わせられているかを解析することによって、高レベルの構造を抽出することができる。

【０２５７】また、本発明にかかる映像音声処理装置は、供給されたビデオ信号の内容の意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置であって、ビデオ信号を構成する連続した映像及び／又は音声フレームのひと続きから形成される映像及び／又は音声セグメントから、その特徴を表す少なくとも１つ以上の特徴量を抽出する特徴量抽出手段と、特徴量を用いて、特徴量のそれぞれ毎に、映像及び／又は音声セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により映像及び／又は音声セグメントの対の間の類似性を測定する類似性測定手段と、特徴量と測定基準とを用いて、映像及び／又は音声セグメントのうち、互いに類似する複数の映像及び／又は音声セグメントから構成される類似チェーンを検出する検出手段とを備える。

【０２５８】したがって、本発明にかかる映像音声処理装置は、ビデオ信号において類似した映像及び／又は音声セグメントの基本的な構造パターンを決定して出力することが可能であり、これらの構造パターンがどのように組み合わせられているかを解析することによって、高レベルのビデオ構造を抽出することが可能となる。

【図面の簡単な説明】

【図１】本発明において適用するビデオデータの構成を説明する図であって、モデル化したビデオデータの構造を説明する図である。

【図２】局所的ビデオ構造を抽出する類似チェーンを説明する図である。

【図３】大局的ビデオ構造を抽出する類似チェーンを説明する図である。

【図４】本発明の実施の形態として示す映像音声処理装置の構成を説明するブロック図である。

【図５】同映像音声処理装置において、ビデオ構造を検出して解析する際の一連の工程を説明するフローチャートである。

【図６】同映像音声処理装置における動的特徴量サンプリング処理を説明する図である。

【図７】基本類似チェーンを説明する図である。

【図８】リンク類似チェーンを説明する図である。

【図９】周期的チェーンを説明する図である。

【図１０】同映像音声処理装置において、バッチクラスタリング技術を用いて基本類似チェーンを検出する際の一連の工程を説明するフローチャートである。

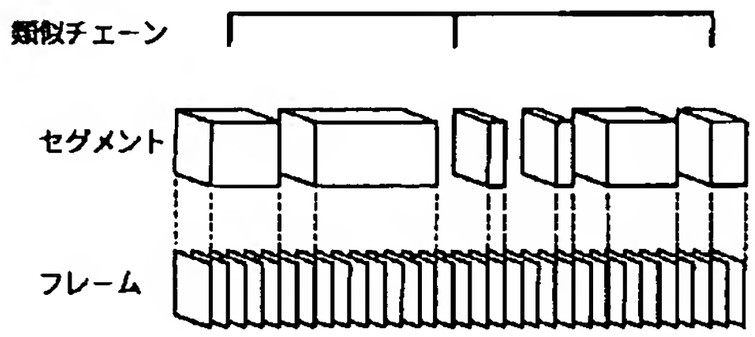
【図１１】非類似性閾値を説明する図である。

【図１２】同映像音声処理装置において、基本類似チェーンのチェーンフィルタリングを行う際の一連の工程を説明するフローチャートである。

【図１３】同映像音声処理装置において、逐次クラスタリング技術を用いて基本類似チェーンを検出する際の一

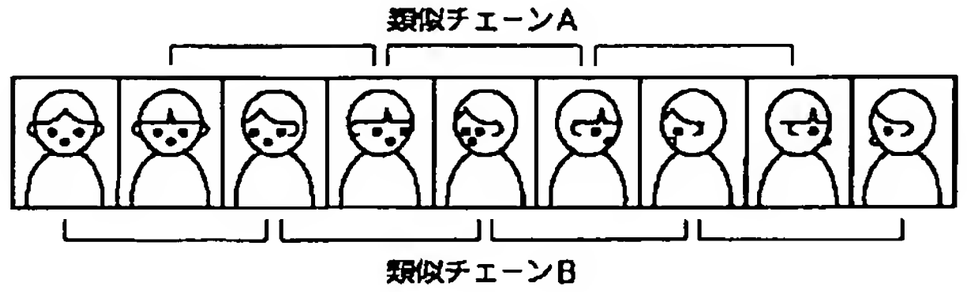
連の工程を説明するフローチャートである。
【図１４】同映像音声処理装置において、リンク類似チェーンを検出する際の一連の工程を説明するフローチャートである。
【図１５】同映像音声処理装置において、周期的チェーンを検出する際の一連の工程を説明するフローチャートである。
【図１６】同映像音声処理装置において、チェーンを用いてシーンを検出する際の一連の工程を説明するフローチャートである。
【図１７】同映像音声処理装置において、チェーンを用いてニュース項目を検出する際の一連の工程を説明するフローチャートである。
【図１８】同映像音声処理装置において、チェーンを用いてスポーツ放送におけるプレイを検出する際の一連の工程を説明するフローチャートである。
【図１９】同映像音声処理装置において、チェーンを用いて周期検出とシーン検出とを組み合わせたトピック検出を行う際の一連の工程を説明するフローチャートである。
【符号の説明】
１０　映像音声処理装置、　１１　ビデオ分割部、　１２　ビデオセグメントメモリ、　１３　映像特徴量抽出部、　１４　音声特徴量抽出部、　１５　セグメント特徴量メモリ、　１６　チェーン検出部、　１７　特徴量類似性測定部、１８　チェーン解析部

【図１】



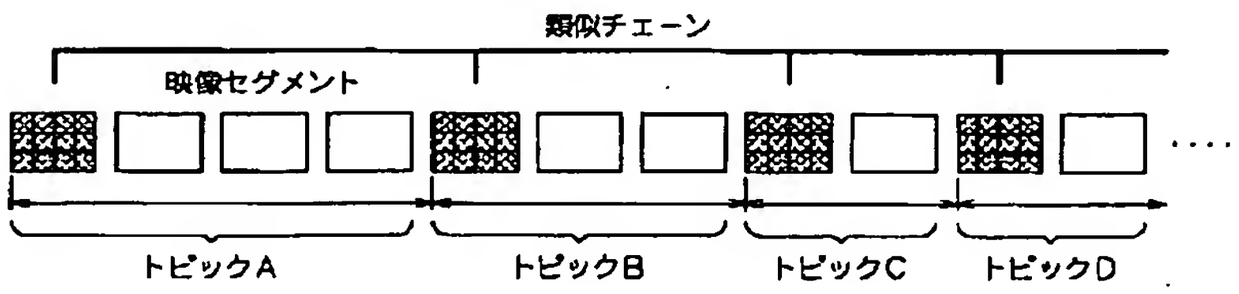
ビデオ構造の階層モデル

【図２】



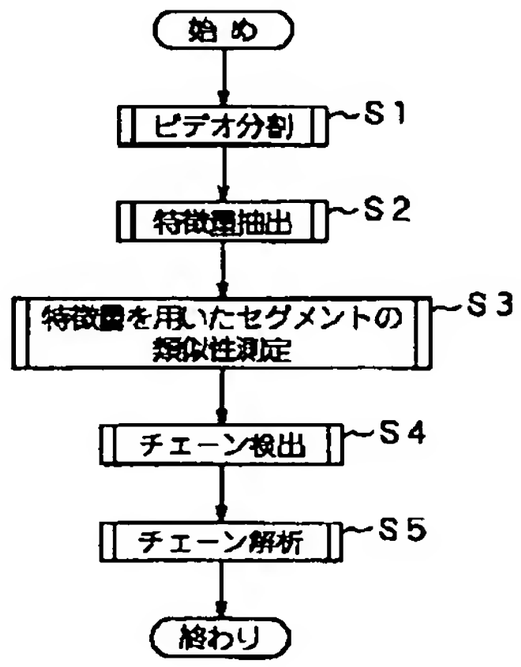
類似チェーンの説明図

【図３】



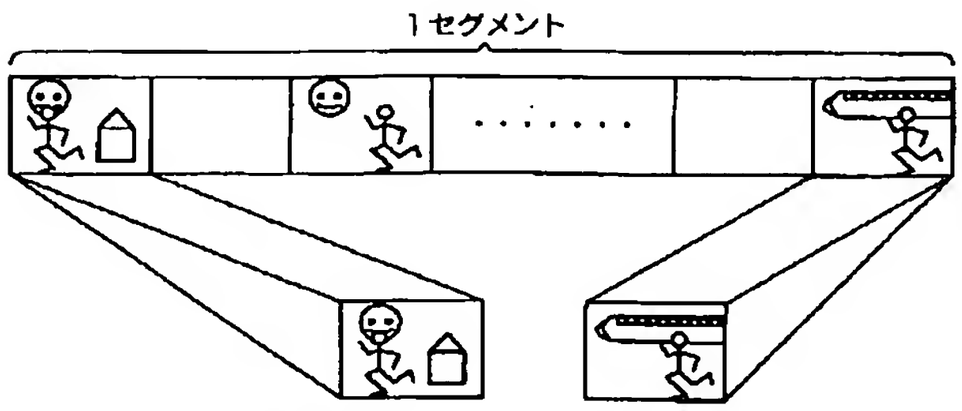
類似チェーンの説明図

【図５】



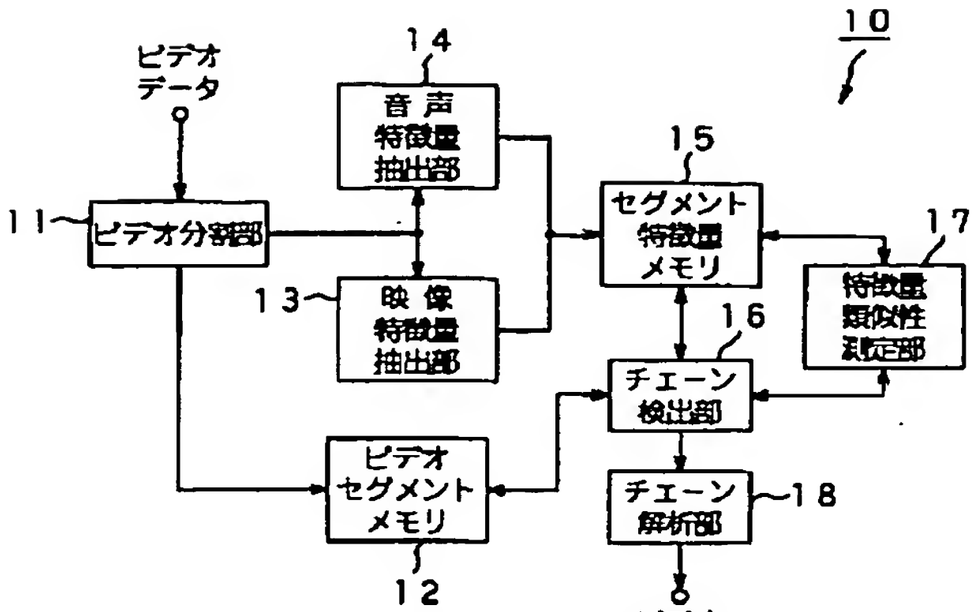
映像音声処理装置における一連の処理工程

【図６】



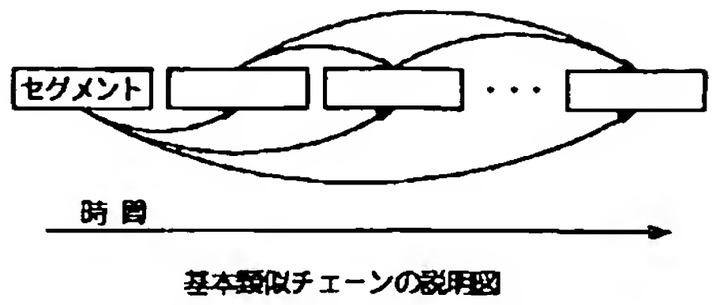
特徴量のサンプリング方法の説明図

【図４】



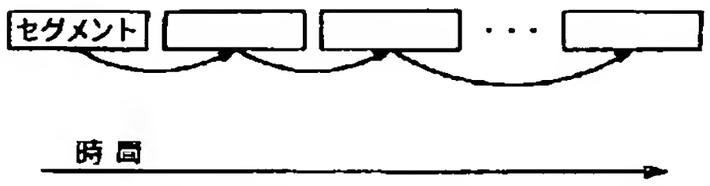
映像音声処理装置の構成ブロック図

【図７】



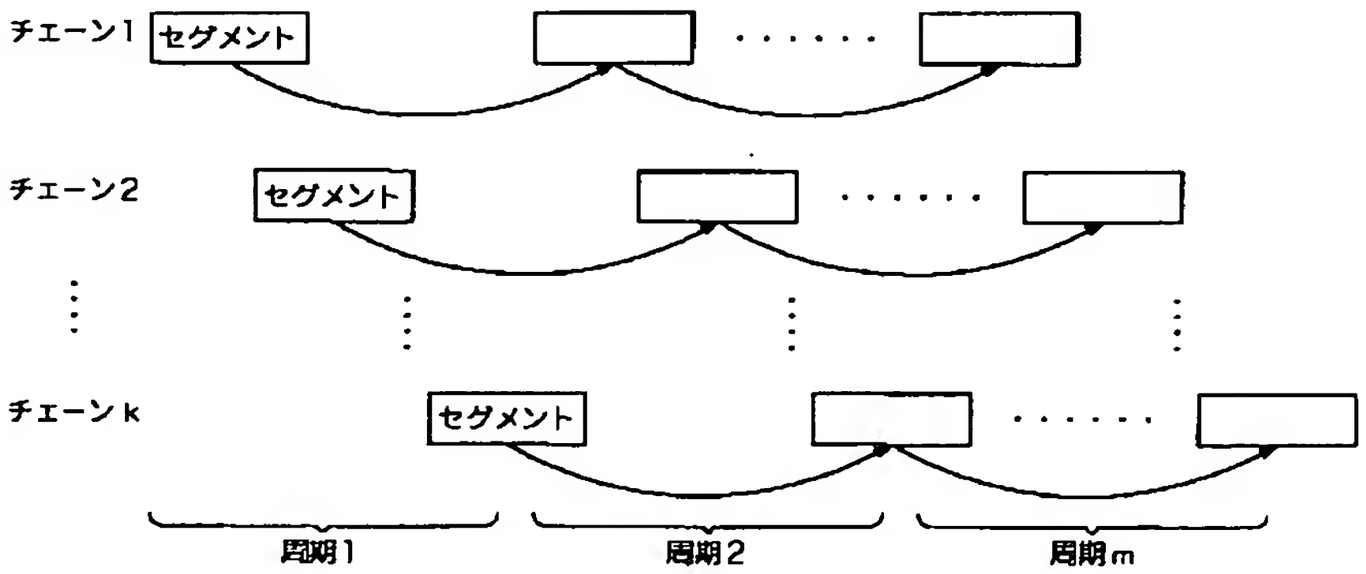
基本類似チェーンの説明図

【図８】



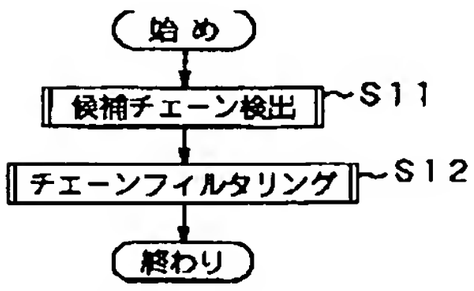
リンク類似チェーンの説明図

【図９】



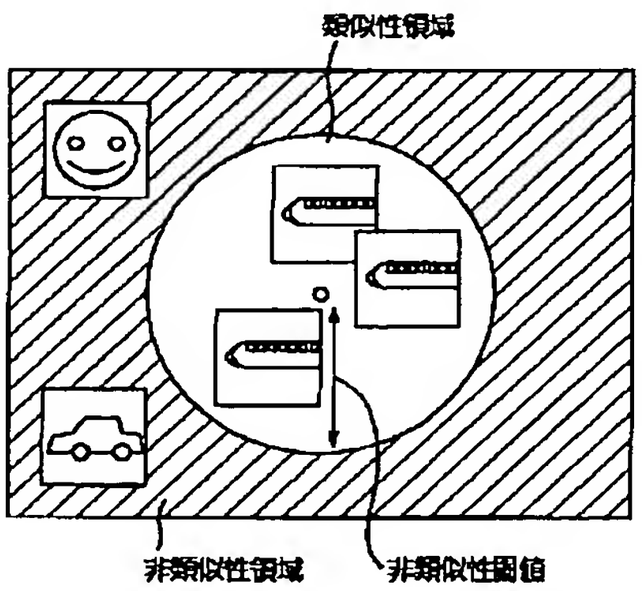
周期的チェーンの説明図

【図１０】



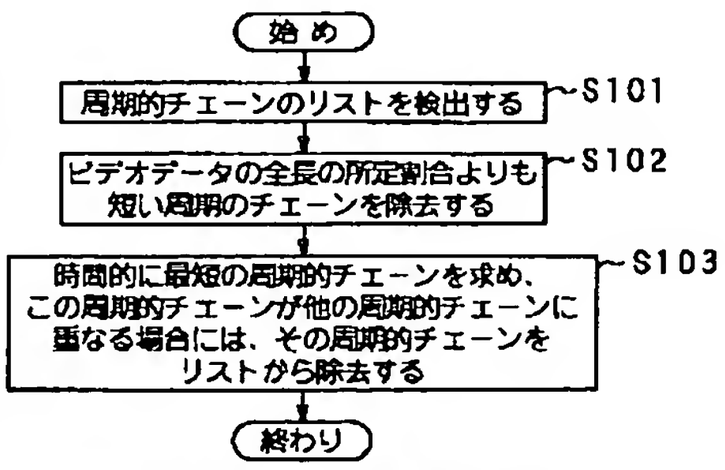
映像音声処理装置における一連の処理工程

【図１１】



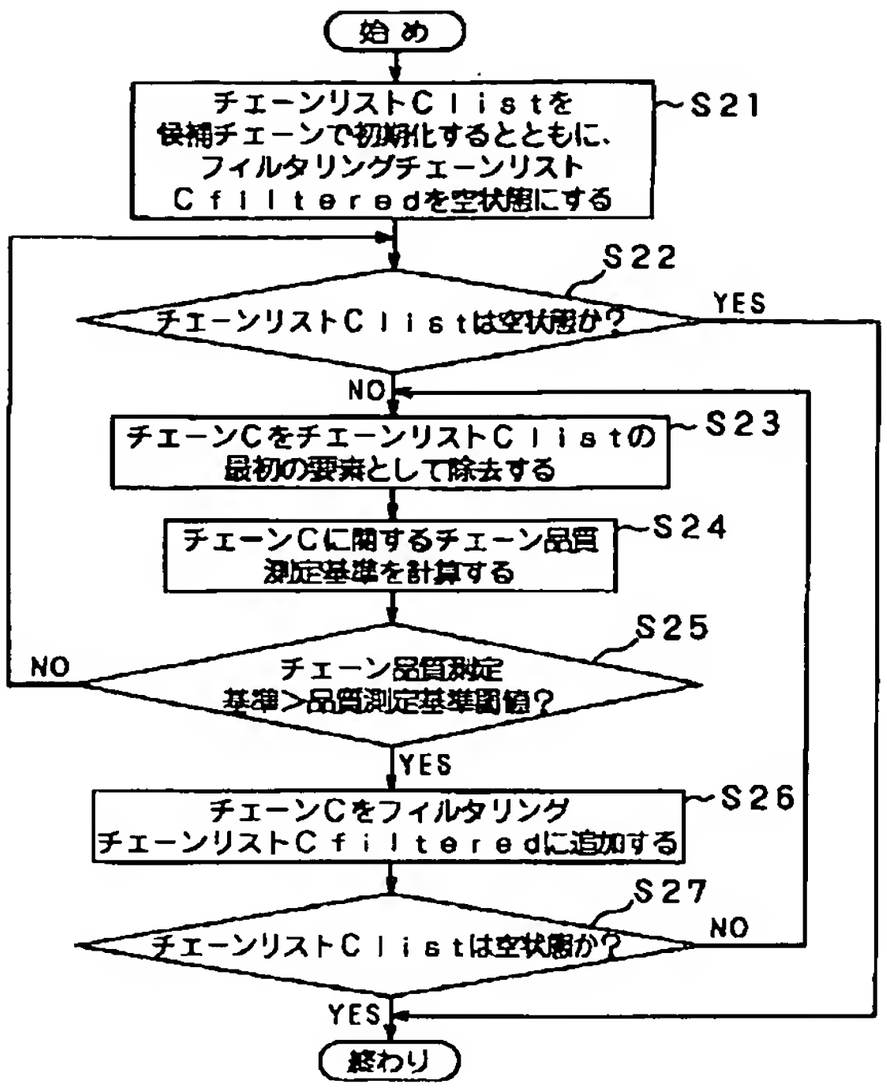
非類似性閾値の説明図

【図１７】



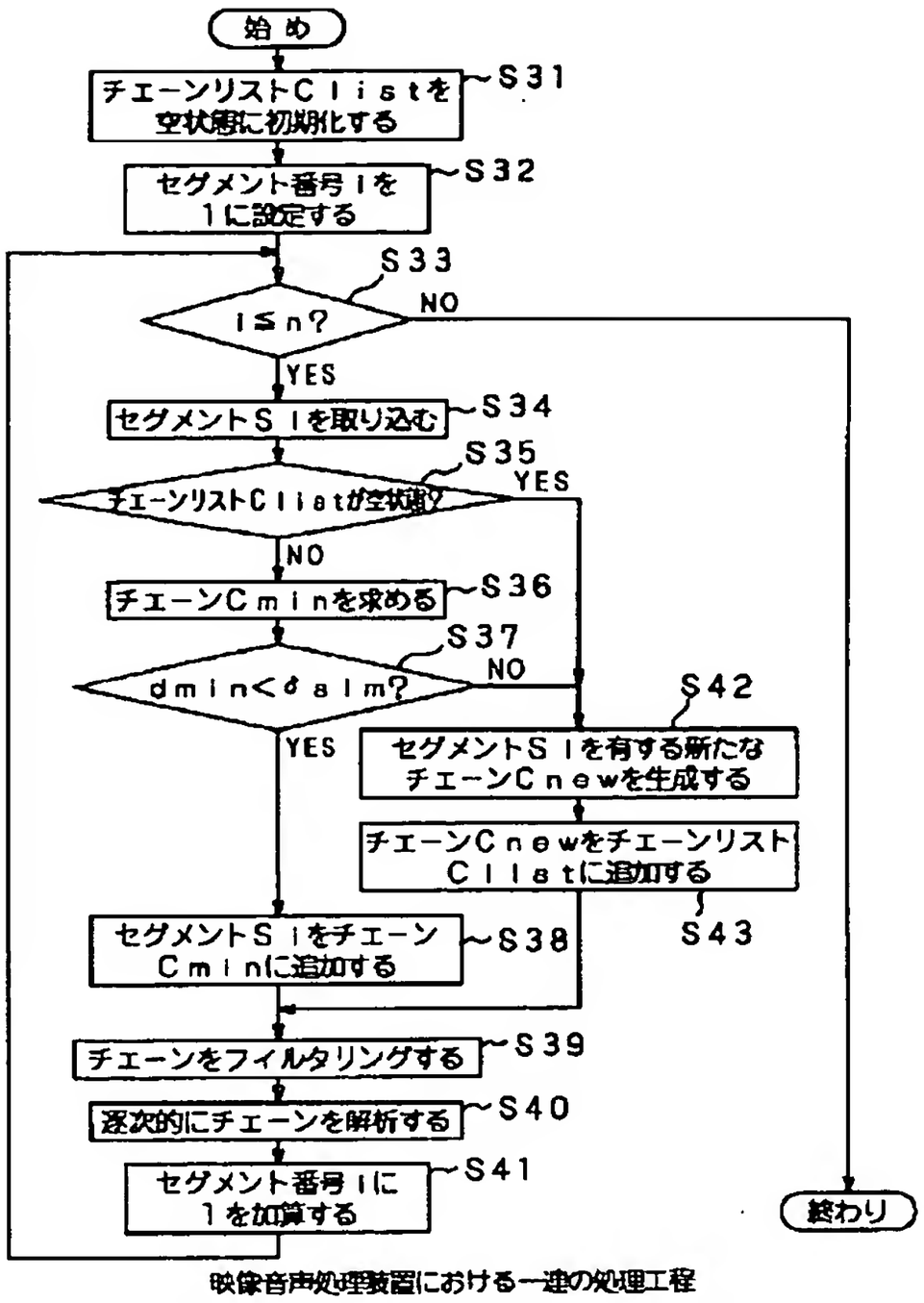
映像音声処理装置における一連の処理工程

【図１２】



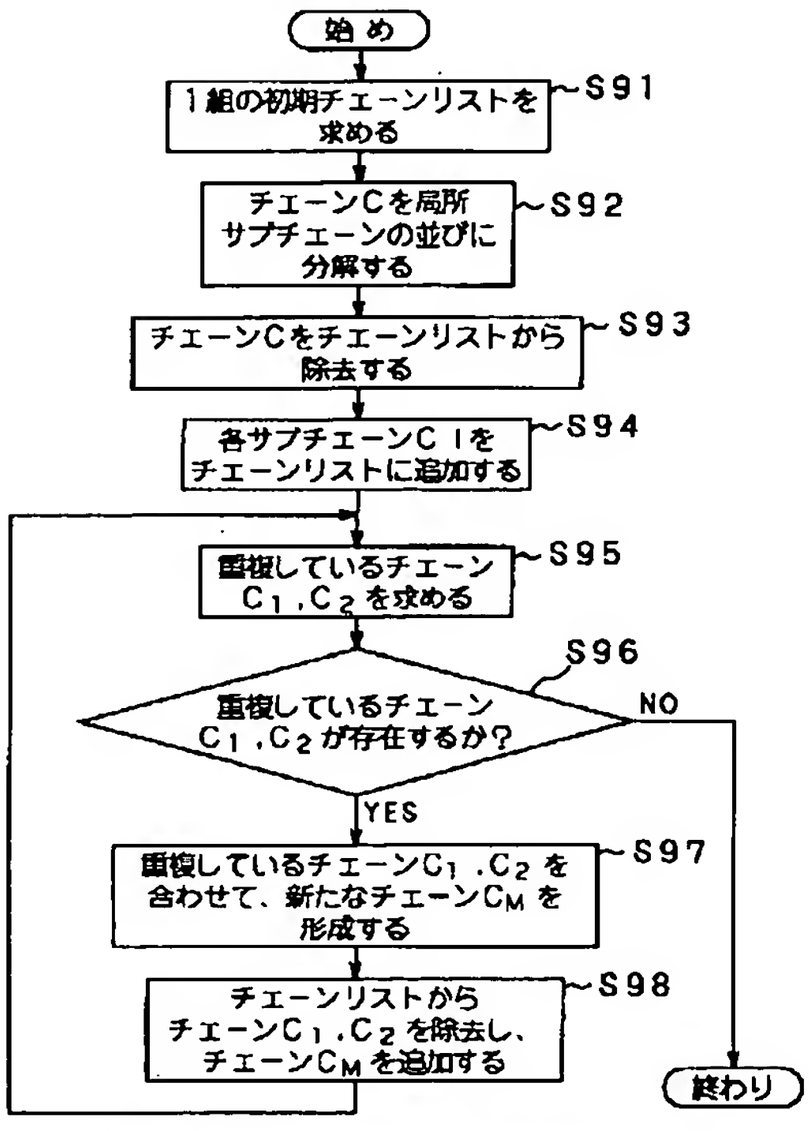
映像音声処理装置における一連の処理工程

【図１３】



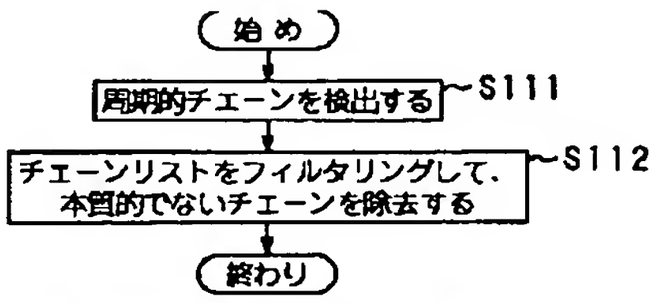
映像音声処理装置における一連の処理工程

【図１６】



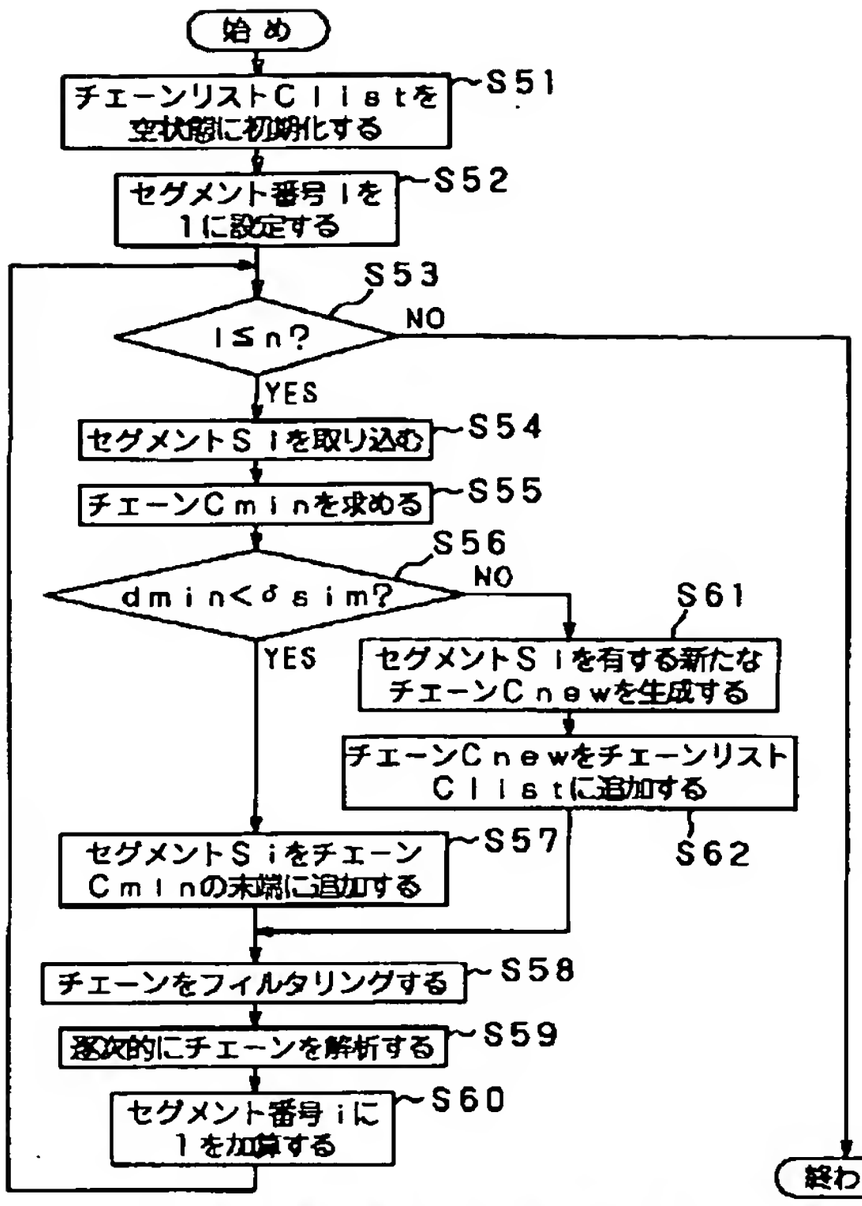
映像音声処理装置における一連の処理工程

【図１８】



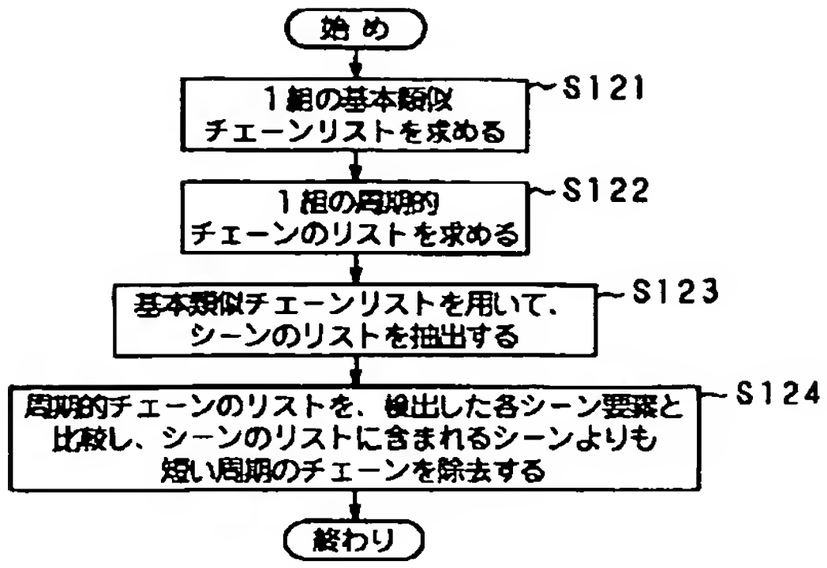
映像音声処理装置における一連の処理工程

【図１４】



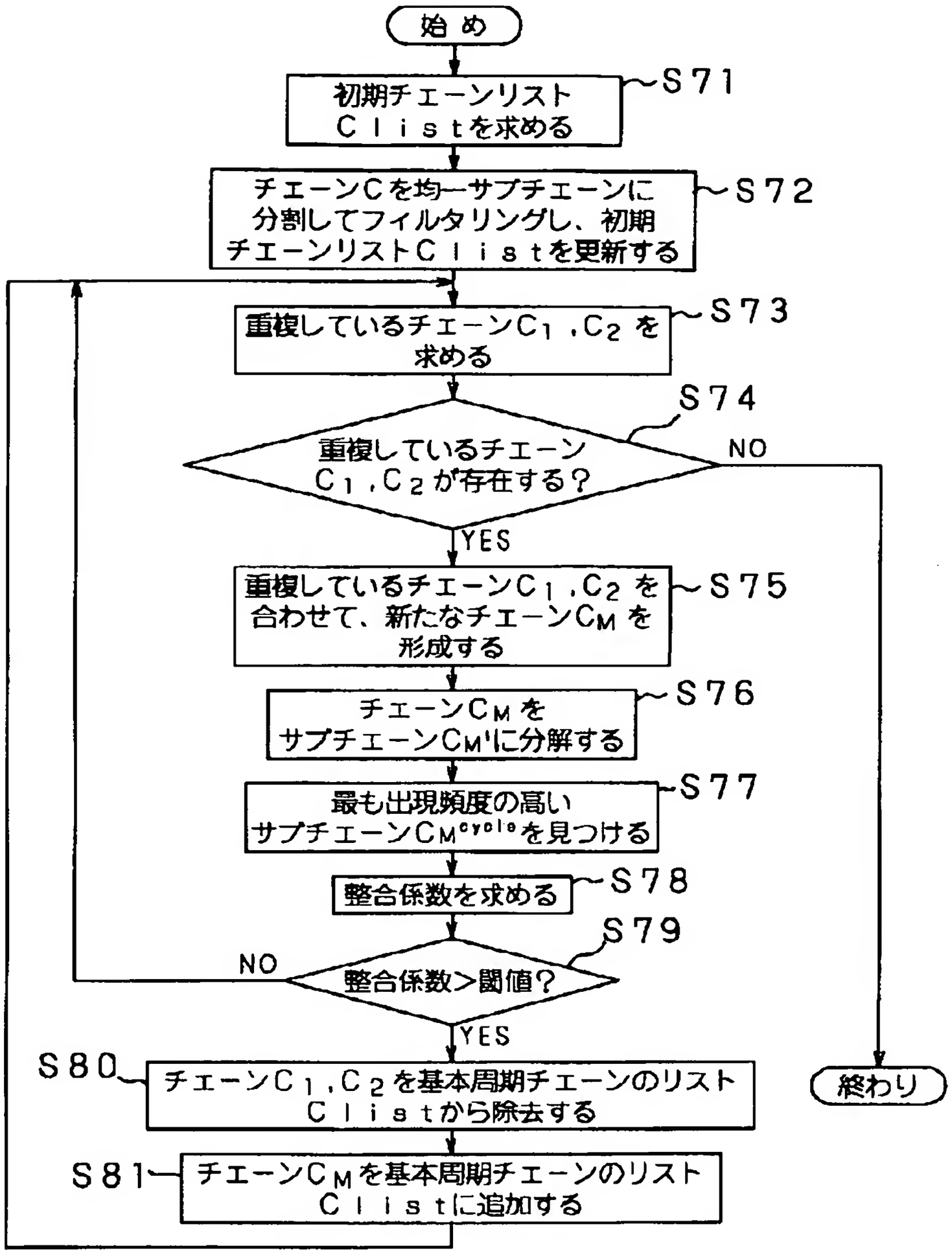
映像音声処理装置における一連の処理工程

【図１９】



映像音声処理装置における一連の処理工程

【図１５】



映像音声処理装置における一連の処理工程

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-285243

(43)Date of publication of application : 13.10.2000

---

(51)Int.Cl. G06T 7/00
H04N 5/91
// G10L 15/00

---

(21)Application number : 2000-023339 (71)Applicant : SONY CORP

(22)Date of filing : 27.01.2000 (72)Inventor : TOBY WALKER

---

(30)Priority
Priority number : 11023069
Priority date : 29.01.1999
Priority country : JP

---

(54) SIGNAL PROCESSING METHOD AND VIDEO SOUND PROCESSING DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To extract a high level video structure in various videos. SOLUTION: A video sound processing device 10 is provided with a chain detection part 16 and a chain analysis part 18. In the chain detection part 16, feature quantities extracted from video segments and/or audio segments divided from a stream of

inputted video data and a measurement reference which is calculated for each feature quantity by using the above feature quantities and measures the similarities between video segments and/or audio segments are used to detect a similar chain consisting of a plurality of video and/or audio segments similar to each other out of video segments and/or audio segments. In the chain analysis part 18, the similar chain is used to perform analysis, and a local video structure and/or a global video structure of the video is determined and outputted.

CLAIMS

[Claim(s)]
[Claim 1]A signal processing method comprising:
A characteristic quantity extraction process of extracting at least one or more characteristic quantity showing the feature from a segment which is a signal processing method which detects and analyzes a pattern reflecting a semantic structure of the contents of the supplied signal, and is formed from a series of a continuous frame which constitutes the above-mentioned signal.
A similarity measuring process which computes metrics which measure similarity between pairs of the above-mentioned segment for every each of the above-mentioned characteristic quantity using the above-mentioned characteristic quantity, and measures similarity between pairs of the above-mentioned segment by these metrics.
A detection process which detects a similar chain which comprises two or more segments mutually similar among the above-mentioned segments using the above-mentioned characteristic quantity and the above-mentioned metrics.

[Claim 2]The signal processing method according to claim 1 provided with an analysis process which analyzes using the above-mentioned similar chain, and determines and

outputs local structure and/or global structure of the above-mentioned signal.

[Claim 3]The signal processing method according to claim 1, wherein the above-mentioned signal is at least one of the video signals and audio signals in a video data.

[Claim 4]The signal processing method according to claim 1, wherein the above-mentioned similar chain has restrictions in a relation between similar segments which the similar chain concerned contains.

[Claim 5]The signal processing method according to claim 1, wherein the above-mentioned similar chain has restrictions in structure of the similar chain concerned.

[Claim 6]The signal processing method according to claim 4, wherein the above-mentioned similar chain is a basic similar chain where all the segments which the similar chain concerned contains have a mutually similar relation.

[Claim 7]The signal processing method according to claim 4, wherein the above-mentioned similar chain is a link similar chain where an adjoining segment has a mutually similar relation in all the segments which the similar chain concerned contains.

[Claim 8]The signal processing method according to claim 4 characterized by being a periodic chain which has a relation similar as mutually [the number of predetermined / from the segment concerned / in each of a segment] as a segment by which it has been arranged back in all the segments in which the similar chain concerned contains the above-mentioned similar chain.

[Claim 9]The signal processing method according to claim 5, wherein the above-mentioned similar chain is a partial chain whose time interval in each set of a segment which adjoins in all the segments which the similar chain concerned contains is shorter than predetermined time.

[Claim 10]The signal processing method according to claim 5, wherein the above-mentioned similar chain is a uniform chain with which a segment appears at intervals of isochronous approximately in all the segments which the similar chain concerned contains.

[Claim 11]The signal processing method comprising according to claim 6:

A candidate chain detection process which the above-mentioned detection process detects and summarizes a mutually similar segment using the above-mentioned characteristic quantity and the above-mentioned metrics, and forms a candidate chain.

A filtering process of outputting only a candidate chain which computes quality metrics corresponding to a numerical standard for every each of the above-mentioned candidate chain, and measures importance and relevance of the above-mentioned candidate chain in structural-patterns analysis of the above-mentioned signal and with which the above-mentioned quality metrics exceed

a predetermined quality metrics threshold.

[Claim 12]The signal processing method according to claim 2 carrying out sequential processing of every one segment concerned according to time order to which a segment was supplied among segments in the above-mentioned signal.
[Claim 13]The signal processing method comprising according to claim 12:
A candidate chain detection process which the above-mentioned detection process updates as required a candidate chain which contains the segment concerned using target above-mentioned characteristic quantity and above-mentioned metrics about a segment, and is searched for.
A filtering process of outputting only a candidate chain which computes quality metrics corresponding to a numerical standard for every each of the above-mentioned candidate chain, and measures importance and relevance of the above-mentioned candidate chain in structural-patterns analysis of the above-mentioned signal and with which the above-mentioned quality metrics exceed a predetermined quality metrics threshold.

[Claim 14]The signal processing method comprising according to claim 8:
An initial periodic chain detection process which the above-mentioned detection process asks for an initial candidate of a periodic chain.
A duplication chain detection process which asks for a duplication chain which crosses in time out of an initial candidate of the above-mentioned periodic chain.
A consistency process of searching for consistency of the above-mentioned duplication chain.

[Claim 15]The signal processing method according to claim 2 detecting and outputting a scene which is a subset based on a meaning of a segment as a local structure of the above-mentioned signal according to the above-mentioned analysis process using the above-mentioned similar chain.
[Claim 16]The signal processing method according to claim 2, wherein a mutually similar segment detects and outputs structural patterns by which it is generated repetitively as a global structure of the above-mentioned signal according to the above-mentioned analysis process using the above-mentioned similar chain.
[Claim 17]The signal processing method according to claim 16 detecting and outputting a news item in newscasting as the above-mentioned structural patterns.
[Claim 18]The signal processing method according to claim 16, wherein a play detects and outputs video structure in a sportscast generated repetitively as the above-mentioned structural patterns.
[Claim 19]The signal processing method according to claim 2 detecting and outputting topic structure which summarized a scene related among scenes which are the

subsets based on a meaning of a segment according to the above-mentioned analysis process using the above-mentioned similar chain.

[Claim 20]A video voice processing unit comprising:

It is a video voice processing unit which detects and analyzes a pattern of an image reflecting a semantic structure of the contents of the supplied video signal, and/or a sound, A feature amount extracting means which extracts at least one or more characteristic quantity showing the feature from an image formed from a series of a continuous image which constitutes the above-mentioned video signal, and/or an audio frame, and/or a sound segment.

A similarity measuring means which computes metrics which measure similarity between pairs of the above-mentioned image and/or a sound segment for every each of the above-mentioned characteristic quantity using the above-mentioned characteristic quantity, and measures similarity between pairs of the above-mentioned image and/or a sound segment by these metrics.

A detection means to detect a similar chain which comprises two or more images and/or sound segments mutually similar among the above-mentioned image and/or a sound segment using the above-mentioned characteristic quantity and the above-mentioned metrics.

[Claim 21]The video voice processing unit according to claim 20 provided with an analysis means to analyze using the above-mentioned similar chain, and to determine and output local video structure and/or global video structure of the above-mentioned video signal.

[Claim 22]The video voice processing unit according to claim 20, wherein the above-mentioned similar chain has restrictions in a relation between a similar image and/or a sound segment which the similar chain concerned contains.

[Claim 23]The video voice processing unit according to claim 20, wherein the above-mentioned similar chain has restrictions in structure of the similar chain concerned.

[Claim 24]The video voice processing unit according to claim 22, wherein the above-mentioned similar chain is a basic similar chain where all the images and/or sound segments which the similar chain concerned contains have a mutually similar relation.

[Claim 25]The video voice processing unit according to claim 22, wherein the above-mentioned similar chain is a link similar chain where an adjoining image and/or a sound segment have a mutually similar relation in all the images and/or sound segments which the similar chain concerned contains.

[Claim 26]In all the images and/or sound segments in which the similar chain concerned contains the above-mentioned similar chain, The video voice processing unit according to claim 22 with which each of an image and/or a sound segment is

characterized by being an image and/or a sound segment by which only a predetermined number has been arranged back, and a periodic chain which has a mutually similar relation from the segment concerned.

[Claim 27]The video voice processing unit according to claim 23, wherein the above-mentioned similar chain is a partial chain whose time interval in an image which adjoins in all the images and/or sound segments which the similar chain concerned contains, and/or each set of a sound segment is shorter than predetermined time.

[Claim 28]The video voice processing unit according to claim 23, wherein the above-mentioned similar chain is a uniform chain with which an image and/or a sound segment appear at intervals of isochronous approximately in all the images and/or sound segments which the similar chain concerned contains.

[Claim 29]The above-mentioned detection means detects and summarizes a mutually similar image and/or a sound segment using the above-mentioned characteristic quantity and the above-mentioned metrics, and forms a candidate chain, Quality metrics corresponding to a numerical standard are computed for every each of the above-mentioned candidate chain, The video voice processing unit according to claim 24, wherein it measures importance and relevance of the above-mentioned candidate chain over structural-patterns analysis of the above-mentioned video signal and the above-mentioned quality metrics output only a candidate chain which exceeds a predetermined quality metrics threshold.

[Claim 30]The video voice processing unit according to claim 21 characterized by carrying out sequential processing of image concerned and/or every one sound segment according to time order to which an image and/or a sound segment were supplied among an image in the above-mentioned video signal, and/or a sound segment.

[Claim 31]The above-mentioned characteristic quantity and the above-mentioned metrics about target above-mentioned present image and/or sound segment are used for the above-mentioned detection means, Update as required a candidate chain containing image concerned and/or a sound segment, ask for it, and quality metrics corresponding to a numerical standard are computed for every each of the above-mentioned candidate chain, The video voice processing unit according to claim 30, wherein it measures importance and relevance of the above-mentioned candidate chain in structural-patterns analysis of the above-mentioned video signal and the above-mentioned quality metrics output only a candidate chain which exceeds a predetermined quality metrics threshold.

[Claim 32]The video voice processing unit according to claim 26, wherein the above-mentioned detection means asks for an initial candidate of a periodic chain, asks for a duplication chain which crosses in time and searches for consistency of the above-mentioned duplication chain out of an initial candidate of the above-mentioned periodic chain.

[Claim 33]The video voice processing unit according to claim 21, wherein the above-mentioned analysis means detects and outputs a scene which are an image and/or a subset based on a meaning of a sound segment as local video structure of the above-mentioned video signal using the above-mentioned similar chain.

[Claim 34]The video voice processing unit according to claim 21, wherein the above-mentioned analysis means detects and outputs structural patterns which a mutually similar image and/or a sound segment generate repetitively as global video structure of the above-mentioned video signal using the above-mentioned similar chain.

[Claim 35]The video voice processing unit according to claim 34, wherein the above-mentioned analysis means detects and outputs a news item in newscasting as the above-mentioned structural patterns.

[Claim 36]The video voice processing unit according to claim 34, wherein the above-mentioned analysis means detects and outputs video structure in a sportscast which a play generates repetitively as the above-mentioned structural patterns.

[Claim 37]The video voice processing unit according to claim 21, wherein the above-mentioned analysis means detects and outputs topic structure which summarized a scene related among scenes which are an image and/or a subset based on a meaning of a sound segment using the above-mentioned similar chain.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention]This invention relates to the video voice processing unit which detects and analyzes the pattern of the image reflecting the semantic structure used as the signal processing method which detects and analyzes the pattern reflecting the semantic structure used as the foundation of a signal, and the foundation of a video signal, and/or a sound.

[0002]

[Description of the Prior Art]For example, there is a case where he would like to play in search of the portion of a request of an interested portion etc. out of the video application constituted with a lot of different picture image data called the TV program recorded on the video data.

[0003]thus, there is a storyboard which is the panel which put in order a series of images describing the major scene of application as general art for extracting desired image contents, and was created. This storyboard decomposes a video data into what is called a shot, and displays the image represented in each shot. Such image

extraction art the most, For example, "G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. ofVisual Communicatio n Detect a shot automatically and extract it from a video data as indicated to and Image Representation 7:28-4 and 1996."

[0004]

[Problem(s) to be Solved by the Invention]By the way, hundreds of shots are also contained, for example in the TV program for 30 typical minutes. Therefore, in the conventional image extraction art mentioned above, when the user needed to investigate the storyboard which put an extracted huge number of shots in order and understood such a storyboard, he needed to force the big burden upon the user. in the conventional image extraction art, the shot in the conversation scene which photoed two persons by turns, for example according to a speaker's change had the problem that there were many redundant things. Thus, as an object which extracts video structure, the hierarchy of the shot was too low, there was much useless amount of information, and the conventional image extraction art of extracting such a shot was not able to be said as the good thing of convenience for the user.

[0005]As other image extraction art, For example, "A. Merlino and D. Morey. and M. Maybury and Broadcast. As indicated to news navigation using story segmentation, Proc. of ACMMultimedia 97, and 1997" or JP,10-136297,A, There is a thing using the very special knowledge about specific contents genres, such as news and a football game. However, as a result of this conventional image extraction art is not helpful to other genres of what can obtain a good result at all about the target genre and being further limited to a genre, there was a problem of not being easily generalizable.

[0006]As other image extraction art, there are some which extract what is called a story unit as indicated, for example in the U.S. Patent # No. 5,708,767 gazette. However, this conventional image extraction art needed a user's intervention, in order that it might not automate thoroughly and which shot might determine whether to be what shows the same contents. It also had the problem that it was limited only to video information as an applied object while this conventional image extraction art had the complicated calculation which processing takes.

[0007]As other image extraction art, there are some which identify a shot by combining shot detection and silent part detection further again as indicated, for example to JP,9-214879,A. However, this conventional image extraction art was limited only when a silent part corresponded to a shot boundary.

[0008]As other image extraction art, For example, "H. Aoki, S. Shimotsuji and O.Hori, A shot classification method to select effective key-frames for video browsing,. In order to reduce the redundancy of the display in a storyboard as indicated to IPSJ Human Interface SIG Notes, 7:43-50, 1996", or JP,9-93588,A, there are some which detect the repeated similar shot. However, this conventional image extraction art can be applied only to video information, and cannot be applied to speech information.

[0009]Such image extraction art [like] was able to detect only what is called local video structure and the global video structure based on special knowledge.

[0010]This invention is made in view of such the actual condition, and solves the problem of the conventional image extraction art mentioned above, and an object of this invention is to provide the signal processing method and video voice processing unit which extract the video structure of the high level in various video datas.

[0011]

[Means for Solving the Problem]A signal processing method concerning this invention which attains the purpose mentioned above, It is a signal processing method which detects and analyzes a pattern reflecting a semantic structure of the contents of the supplied signal, A characteristic quantity extraction process of extracting at least one or more characteristic quantity showing the feature from a segment formed from a series of a continuous frame which constitutes a signal, Metrics which measure similarity between pairs of a segment are computed for every each of characteristic quantity using characteristic quantity, It is characterized by having a detection process which detects a similar chain which comprises two or more segments mutually similar among segments using a similarity measuring process which measures similarity between pairs of a segment by these metrics, and characteristic quantity and metrics.

[0012]A signal processing method concerning such this invention detects fundamental structural patterns of a segment similar in a signal.

[0013]A video voice processing unit concerning this invention which attains the purpose mentioned above, It is a video voice processing unit which detects and analyzes a pattern of an image reflecting a semantic structure of the contents of the supplied video signal, and/or a sound, A feature amount extracting means which extracts at least one or more characteristic quantity showing the feature from an image formed from a series of a continuous image which constitutes a video signal, and/or an audio frame, and/or a sound segment, Metrics which measure similarity between pairs of an image and/or a sound segment are computed for every each of characteristic quantity using characteristic quantity, A similarity measuring means which measures similarity between pairs of an image and/or a sound segment by these metrics, It is characterized by having a detection means to detect a similar chain which comprises two or more images and/or sound segments mutually similar among an image and/or a sound segment, using characteristic quantity and metrics.

[0014]A video voice processing unit concerning such this invention determines and outputs an image similar in a video signal, and/or fundamental structural patterns of a sound segment.

[0015]

[Embodiment of the Invention]It explains in detail, referring to drawings for the concrete embodiment which applied this invention hereafter.

[0016]The embodiment which applied this invention is a video voice processing unit which discovers the desired contents automatically and extracts them from the recorded video data. Especially this video voice processing unit introduces the concept of a similar chain (it is hereafter written as a chain if needed.), in order to detect and analyze the structural patterns of the image reflecting the semantic structure used as the foundation of a video data, and/or a sound and to conduct this analysis. Before giving concrete explanation of this video voice processing unit, in this invention, explanation about the target video data is given here first.

[0017]In this invention, about the target video data, as shown in drawing 1, a model shall be made, and it shall have structure of a frame, a segment, and a similar chain. That is, a video data is constituted by a series of frames in a least significant layer. A video data is constituted by the segment formed from a series of a continuous frame as a hierarchy on one of the frames.A video data constitutes a series of segments which have a specific kind of similar pattern of each other as a similar chain.

[0018]This video data includes the information on both an image and a sound. That is, in this video data, the video frame which is a single still picture, and the audio frame showing the speech information by which the sample was generally carried out in short time, such as tens – hundreds milliseconds / merit, are contained in a frame.

[0019]A segment comprises a series of the video frame continuously photoed with the single camera, and, generally is called a shot. And video segments and a sound segment are contained in a segment, and it becomes a basic unit in video structure. In these segments, many definitions are possible about especially a sound segment, and a thing as shown below as an example can be considered. First, the sound segment can appoint a boundary by the silent period in the video data detected by the method generally known well, and may be formed. A sound segment as indicated to "D.Kimber and L. Wilcox, Acoustic Segmentation for Audio Browsers, and Xerox Parc Technical Report", For example, it may be formed from a series of a sound, music, a noise, and the audio frame classified into a small number of category like silent **. A sound segment, "S. Pfeiffer,S. Fischer and E. Wolfgang, Automatic Audio Content Analysis, Proceeding of ACMMultimedia 96, Nov. 1996, The big change in a certain feature between two continuous audio frames is detected as a voice cut point, and it may be determined based on this voice cut point as indicated to pp21–30."

[0020]In such a video data, with a similar chain. It is mutually similar, and it is two or more segments which were able to be set in order in time, and the structural patterns are classified into some kinds according to the constraints which should be fulfilled as the relation between the similar segments contained in the chain concerned, and a structure of a chain. Formally, similar chains are a series of segments of which $j= 1, ..., k-1: i_j < i_{j+1}$ consists about all the segments, when the segment which the similar chain concerned contains is expressed with $Si_1, ..., Si_k$. Index i_j expresses the segment number in the video data of the origin of the segment here, and subscript j to i means

that the segment is located on a time-axis in the similar chain concerned the j-th. Since a discontinuous segment is contained in a similar chain in time, a time gap may exist between the elements of a chain. If it puts in another way, segment S_{ij} and S_{ij+1} will not necessarily continue in the original video data.

[0021]By using a similar chain, the leading key about both the local video structures and global video structures which are mentioned later can be acquired in a video data. Generally the key as which a televiewer can grasp the outline perceptually exists in a video data. Things simplest as this key and important are similar structural patterns of video segments or a sound segment, and are the information which should gain just these structural patterns with a similar chain.

[0022]As such a similar chain, there are a basic similar chain, a link similar chain, a partial chain, and a periodic chain, and these are the most important and fundamental in video-data analysis so that it may explain in full detail behind.

[0023]Here, with a basic similar chain, all the segments which the basic similar chain concerned contains are mutually similar. However, there are no restrictions in the structural patterns. Generally such a basic similar chain can be obtained using the grouping algorithm or clustering algorithm for carrying out grouping of the segment. With a link similar chain, the segment which adjoins in the chain is mutually similar. A partial chain has a time interval smaller than predetermined time between segments in each set of the adjoining segment. And each segment is [chain / periodic] similar with the segment of the m-th back rather than it. That is, a periodic chain comprises that m segments are repeated approximately.

[0024]And such a similar chain can be used for extracting the local video structure of a scene, for example, the global video structure of a news item, also in a video data as shown below.

[0025]In order to describe a video data to be a scene on a higher level based on the semantic content, here, Grouping is carried out to the settlement which is meaningful using the characteristic quantity showing the feature of a segment called the amount of perceptual activities in a segment for example in the segment obtained by video-segments (shot) detection or sound segment detection. Although the scene was subjective and it was dependent on the contents or the genre of a video data, the characteristic quantity should carry out grouping of the repetitive pattern of the video segments which show similarity mutually, or a sound segment here.

[0026]Now, as an example of a similar chain of extracting the local video structure mentioned above, as shown in drawing 2, in the scene where two speakers are talking mutually, video segments consider the case where it appears by turns according to a speaker. In the video data which has such a repetitive pattern, each video segments are constituted for every ingredient of A ingredient and B ingredient by two crossing chains. Therefore, generally such a crossing partial chain can be used for detecting the related group or scene of video segments.

[0027]As an example of a similar chain of extracting the global video structure mentioned above, as shown in drawing 3, the news program which has fixing structure is considered. In such a video data, the segment to which a newscaster introduces an item for every news item appears first, and the segment which a special correspondent reports, for example from a spot appears following it. In the video data which has such fixing structure, the video segments of the newscaster who appears repeatedly constitute a global chain. Here, since a newscaster's segment shows the start part of each news item, it can detect a news item automatically by using a global chain. That is, in the figure, each topic is detectable by using a global chain out of the video data which comprises two or more news items called the topic A, B, and C, D, and ...

[0028]The video voice processing unit 10 shown in drawing 4 as an embodiment which applied this invention, The similar chain which measured and mentioned the similarity between segments above using the characteristic quantity of the segment in the video data mentioned above is detected automatically, and it can apply to both video segments and a sound segment. And the video voice processing unit 10 can extract and reconstruct the structure of high level, such as a scene which is local video structure, and a topic which is global video structure, from a video data by analyzing a similar chain.

[0029]The video voice processing unit 10 is provided with the following.

The video dividing part 11 which divides the stream of the inputted video data into the segment of an image, sounds, or these both as shown in the figure.

The video segment memory 12 which memorizes the partition information of a video data.

The image feature quantity extracting part 13 which is a feature amount extracting means which extracts the characteristic quantity in each video segments.

The voice feature amount extraction part 14 which is a feature amount extracting means which extracts the characteristic quantity in each sound segment, The segment characteristic quantity memory 15 which memorizes the characteristic quantity of video segments and a sound segment, The chain primary detecting element 16 which is a detection means to summarize video segments and a sound segment to a chain, the characteristic quantity similarity test section 17 which is the similarity measuring means which measure the similarity between two segments, and the chain analyzing parts 18 which are analysis means to detect various video structures.

[0030]The video dividing part 11, for example MPEG1 (Moving Picture Experts Group phase 1) and MPEG 2 (Moving Picture Experts Group phase 2), Or the stream of the video data which consists of the picture image data and voice data in the digitized format of versatility including a compression video-data format like what is called DV

(Digital Video) is inputted, This video data is divided into the segment of an image, sounds, or these both. This video dividing part 11 can be processed directly, without carrying out full extension of this compressed video data, when the inputted video data is a compression format. The video dividing part 11 processes the inputted video data, and divides it into video segments and a sound segment. The video dividing part 11 supplies the partition information which is the result of dividing the inputted video data to the latter video segment memory 12. The video dividing part 11 supplies partition information to the latter image feature quantity extracting part 13 and the voice feature amount extraction part 14 according to video segments and a sound segment.

[0031]The video segment memory 12 memorizes the partition information of the video data supplied from the video dividing part 11. The video segment memory 12 supplies partition information to the chain primary detecting element 16 according to the inquiry from the chain primary detecting element 16 mentioned later.

[0032]The image feature quantity extracting part 13 extracts the characteristic quantity for every video segments obtained by dividing a video data by the video dividing part 11. The image feature quantity extracting part 13 can be processed directly, without carrying out full extension of the compression video data. The image feature quantity extracting part 13 supplies the characteristic quantity of each extracted video segments to the latter segment characteristic quantity memory 15.

[0033]The voice feature amount extraction part 14 extracts the characteristic quantity for every sound segment obtained by dividing a video data by the video dividing part 11. The voice feature amount extraction part 14 can be processed directly, without carrying out full extension of the compression audio data. The voice feature amount extraction part 14 supplies the characteristic quantity of each extracted sound segment to the latter segment characteristic quantity memory 15.

[0034]The segment characteristic quantity memory 15 memorizes the characteristic quantity of the video segments supplied, respectively from the image feature quantity extracting part 13 and the voice feature amount extraction part 14, and a sound segment. The segment characteristic quantity memory 15 supplies the characteristic quantity and the segment which have been memorized to the characteristic quantity similarity test section 17 according to the inquiry from the characteristic quantity similarity test section 17 mentioned later.

[0035]The chain primary detecting element 16 summarizes video segments and a sound segment to a chain using the partition information held at the video segment memory 12, and the similarity between 1 paired segments, respectively. It starts from each segment in a group, and the chain primary detecting element 16 detects the repetitive pattern of a segment similar out of a segment group, and summarizes such a segment to the chain. This chain primary detecting element 16 determines the last set of a chain using the 2nd filtering phase, after summarizing the initial candidate of a

chain. And the chain primary detecting element 16 supplies the detected chain to the latter chain analyzing parts 18.

[0036]The characteristic quantity similarity test section 17 measures the similarity between two segments. The characteristic quantity similarity test section 17 asks the segment characteristic quantity memory 15 that the characteristic quantity about a certain segment is searched.

[0037]The chain analyzing parts 18 analyze the chain structure detected by the chain primary detecting element 16, and detect various local video structures and global video structures. These chain analyzing parts 18 can adjust those details according to specific application so that it may mention later.

[0038]Such a video voice processing unit 10 detects video structure by performing a series of processings in which an outline is shown in drawing 5 using a similar chain.

[0039]First, the video voice processing unit 10 performs video division in Step S1, as shown in the figure. namely, the video data as which the video voice processing unit 10 was inputted into the video dividing part 11 -- either video segments or a sound segment -- or if possible, it will divide into the both. The video voice processing unit 10 does not provide a prerequisite requirement in particular in the video split method to apply. For example, the video voice processing unit 10, "G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication Video division is performed by a method which is indicated to and Image Representation 7:28-4 and 1996." The method of such video division shall be well learned for the technical field concerned, and the video voice processing unit 10 shall apply any video split methods.

[0040]Then, the video voice processing unit 10 extracts characteristic quantity in Step S2. That is, the video voice processing unit 10 calculates the characteristic quantity showing the feature of the segment by the image feature quantity extracting part 13 or the voice feature amount extraction part 14. In the video voice processing unit 10, it is calculated as image characteristic quantity called the time length, color histogram, and texture feature of each segment, voice feature amounts, such as a frequency analysis result, a level, and a pitch, and characteristic quantity which an activity measurement result etc. can apply, for example. Of course, the video voice processing unit 10 is not limited to these as applicable characteristic quantity.

[0041]Then, the video voice processing unit 10 performs similarity measurement of the segment using characteristic quantity in Step S3. That is, the video voice processing unit 10 performs dissimilarity nature measurement by the characteristic quantity similarity test section 17, and measures how many two segments are similar by the metrics. The video voice processing unit 10 calculates dissimilarity nature metrics using the characteristic quantity extracted in previous Step S2.

[0042]Then, the video voice processing unit 10 detects a chain in step S4. That is, the video voice processing unit 10 detects the chain of a similar segment using the

dissimilarity nature metrics calculated in previous Step S3, and the characteristic quantity extracted in previous Step S2.

[0043]And the video voice processing unit 10 analyzes a chain in Step S5. That is, the video voice processing unit 10 determines and outputs the local video structure and/or global video structure of a video data using the chain detected in previous step S4.

[0044]By passing through such a series of processings, the video voice processing unit 10 can detect video structure from a video data. Therefore, a user becomes possible [performing the indexing and abstract of the contents of a video data, or accessing the interested point in a video data promptly] by using this result.

[0045]Hereafter, the processing in the video voice processing unit 10 shown in the figure is explained in detail by every process.

[0046]First, the video division in Step S1 is explained. the video data as which the video voice processing unit 10 was inputted into the video dividing part 11 -- either video segments or a sound segment -- or it dividing into the both, if possible, but. The art for detecting the boundary of the segment in this video data automatically has many things, and it is as having mentioned above in the video voice processing unit 10 concerned not to establish a prerequisite requirement special to this video split method. On the other hand, in the video voice processing unit 10, it depends for the accuracy of the chain detection by a next process on the accuracy of the video division used as the foundation intrinsically.

[0047]Below, the characteristic quantity extraction in Step S2 is explained. Characteristic quantity is the attribute of the segment which supplies the data for measuring the similarity between different segments while expressing the feature of a segment. The video voice processing unit 10 calculates the characteristic quantity of each segment by the image feature quantity extracting part 13 or the voice feature amount extraction part 14, and expresses the feature of a segment. Although it does not depend for the video voice processing unit 10 on the concrete detail of any characteristic quantity, there is a thing like the image characteristic quantity, the voice feature amount, and the amount of video voice common characteristics which are shown below, for example as characteristic quantity which uses in the video voice processing unit 10 concerned, and is considered to be effective. The necessary condition of such characteristic quantity which becomes applicable in the video voice processing unit 10 is that measurement of dissimilarity nature is possible. Such characteristic quantity needs to make it possible to perform simultaneously characteristic quantity extraction and video division mentioned above for increase in efficiency of the video voice processing unit 10. The characteristic quantity explained below fulfills these necessary conditions.

[0048]As characteristic quantity, the thing about an image is mentioned first. Below, this will be called image characteristic quantity. Since video segments are constituted

by the continuous video frame, by extracting a suitable video frame from video segments, they can represent the contents of depiction of the video segments with the extracted video frame, and can express them. That is, the similarity of the video frame extracted appropriately can be substituted for the similarity of video segments. This to image characteristic quantity is one of the important characteristic quantity which can be used with the video voice processing unit 10. If the image characteristic quantity in this case is independent, only static information can be expressed, but the video voice processing unit 10 can also extract the dynamic feature of video segments based on this image characteristic quantity by applying a method which is mentioned later.

[0049]In the video voice processing unit 10, the color in an image serves as an important material at the time of judging whether two images are similar. Judging the similarity of an image using a color histogram, For example, "G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communicati It is well known as indicated to on andImage Representation 7:28-4 and 1996." Here, with a color histogram, three-dimensional color spaces, such as HSV and RGB, are divided into n fields, for example, and the relative ratio of the frequency of occurrence in each field of the pixel in an image is calculated. And a n vector is given from the acquired information. Also about the compressed video data, a color histogram can be extracted directly from compressed data as indicated, for example in the U.S. Patent # No. 5,708,767 gazette.

[0050]In the video voice processing unit 10, the $2^{2 \text{ and } 3}$=64 dimension histogram vector which carried out the sample of the YUV color space from the first in the image which constitutes a segment, and constituted it from 2 bits per color channel is used.

[0051]Although such a histogram expresses the overall color tone of an image, the hour entry is not included in this. Then, in the video voice processing unit 10, imagery correlation is calculated as another image characteristic quantity. In the chain detection in the video voice processing unit 10, the structure which two or more similar segments intersected mutually serves as a leading index which shows that it is one chain structure whose it settled. For example, in a conversation scene, although the position of a camera moves by turns between two speakers, a camera returns to the almost same position, when usually photoing the same speaker again. In such a case, in order to detect the structure where it can set, Since the correlation based on the contraction image of a gray scale found out becoming an index with the good similarity of a segment, in the video voice processing unit 10, the original image is thinned out to the gray scale image of the size of MxN, it reduces, and imagery correlation is calculated using this. Here, a value with small both is enough as M and N, for example, they are 8x8. That is, these reduction gray scale images are interpreted as a feature amount vector of MN dimension.

[0052]The thing about a sound is mentioned as different characteristic quantity from

the image characteristic quantity furthermore mentioned above. Below, this characteristic quantity will be called a voice feature amount. A voice feature amount is the characteristic quantity which can express the contents of the sound segment, and frequency analysis, a pitch, a level, etc. can be used for the video voice processing unit 10 as this voice feature amount. These voice feature amounts are known with various articles.

[0053]First, the video voice processing unit 10 can determine distribution of the frequency information in a single audio frame by conducting frequency analysis, such as the Fourier transform. Since distribution of the frequency information covering one sound segment is expressed, the characteristic quantity of an FFT (Fast Fourier Transform; Fast Fourier Transform) ingredient, a frequency histogram, a power spectrum, and others can be used for the video voice processing unit 10, for example.

[0054]Sound levels, such as pitches, such as an average pitch and a maximum pitch, average loudness, and the maximum loudness, can also use the video voice processing unit 10 as an effective voice feature amount showing a sound segment.

[0055]The video voice processing unit 10 contains a cepstrum coefficient and its primary secondary differential quotient as cepstrum characteristic quantity, The cepstrum spectrum coefficient obtained from an FFT spectrum or LPC (Linear Predictive Coding; linear predictive coding) can also be used.

[0056]As characteristic quantity of further others, the amount of video voice common characteristics is mentioned. Although this is not image characteristic quantity, either and is not a voice feature amount, either, in the video voice processing unit 10, it gives useful information to expressing the feature of the segment in a chain. An activity is used for the video voice processing unit 10 as this amount of video voice common characteristics.

[0057]An activity is an index showing whether the contents of the segment are sensed to be how much dynamic or static. For example, when visually dynamic, an activity expresses the degree from which the degree which a camera moves promptly in accordance with a subject, or the object currently photoed changes promptly.

[0058]This activity is indirectly calculated by measuring the average value of the inter-frame dissimilarity nature of characteristic quantity like a color histogram. Here, if the dissimilarity nature metrics over the characteristic quantity F measured between the frame i and the frame j are defined as d_F (i, j), image activity V_F will be defined like a following formula (1).

[0059]

[Equation 1]

$$V_F = \frac{\sum_{i=b}^{f-1} d_F(i, i+1)}{f-b} \qquad \cdots (1)$$

[0060]In a formula (1), b and f are the frame numbers of a frame of the beginning and the last in one segment, respectively. Specifically, the video voice processing unit 10 can calculate image activity V_F using a histogram mentioned above, for example.

[0061]By the way, fundamentally, although it is as having mentioned above that it is a thing showing static information of a segment, characteristic quantity including image characteristic quantity mentioned above also needs to take dynamic information into consideration, in order to express the feature of a segment correctly. Then, suppose the video voice processing unit 10 that dynamic information is expressed with sampling of characteristic quantity as shown below.

[0062]The video voice processing unit 10 extracts one or more static characteristic quantity from a time of differing in 1 segment, as shown, for example in drawing 6. At this time, the video voice processing unit 10 is determined by balancing maximization of fidelity and minimization of data relative redundancy. [in / for the number of extraction of characteristic quantity / that segment expression] For example, when one certain segment picture can specify as a key-frame of the segment concerned, a histogram calculated from the key-frame serves as sampling characteristic quantity which should be extracted.

[0063]By the way, a certain sample always considers a case where it is chosen, for example at the time of the last in a segment, at the predetermined time. In this case, about two arbitrary segments which change to a black frame (fade), since a sample serves as the same black frame, there is a possibility of bringing a result from which the same characteristic quantity is obtained. That is, image contents of these segments are what kind of things, and it will be judged that that and two selected frames are extremely similar. Since a sample is not a good central value, such a problem is generated.

[0064]Then, suppose the video voice processing unit 10 that characteristic quantity is not extracted in this way in the fixed point, but a statistical central value in the whole segment is extracted. Here, sampling of general characteristic quantity is explained about two cases, i.e., when (1) characteristic quantity can be expressed as a n vector of the real number, and a case where only (2) dissimilarity nature metrics can be used. Image characteristic quantity and voice feature amounts which are known best, such as a histogram and a power spectrum, are contained in (1).

[0065]In (1), a priori, a sample number is decided to be k and the video voice processing unit 10, To "L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990." Characteristic quantity about the whole segment is automatically divided into k different groups using the k average value clustering method (k-means-clustering method) which may be indicated and is known. And the video voice processing unit 10 chooses each group to k groups' centroid value (centroid), or a sample near this centroid value as a sampled value.

Complexity of this processing in the video voice processing unit 10 remains for only increasing linearly about a sample number.

[0066]On the other hand in (2), the video voice processing unit 10, To "L. Kaufman and P.J.Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990." k indicated - a prospect -- ide -- k groups are formed using the algorithm method (k-medoids algorithm method). and a prospect of a group who mentioned above the video voice processing unit 10 for every k groups as a sampled value -- the ide (medoid) is used.

[0067]In the video voice processing unit 10, although based on dissimilarity nature metrics of static characteristic quantity used as the foundation, this is later mentioned for a method of constituting dissimilarity nature metrics about characteristic quantity showing extracted dynamic features.

[0068]Thus, the video voice processing unit 10 can express dynamic features by extracting two or more static characteristic quantity, and using static characteristic quantity of these plurality.

[0069]As mentioned above, the video voice processing unit 10 can extract various characteristic quantity. It is common for each of such characteristic quantity to be insufficient for generally, expressing the feature of a segment, if single. Then, the video voice processing unit 10 can choose a group of characteristic quantity mutually complemented with combining such various characteristic quantity. For example, the video voice processing unit 10 can acquire many information rather than information which each characteristic quantity has by combining a color histogram and imagery correlation which were mentioned above.

[0070]Below, similarity measurement of a segment using characteristic quantity in the step S3 in drawing 5 is explained. The video voice processing unit 10 performs similarity measurement of a segment by the characteristic quantity similarity test section 17 about two characteristic quantity using dissimilarity nature metrics which are a function which calculates the real value which measures how much it is a dissimilarity. When that value of these dissimilarity nature metrics is small, it is shown that two characteristic quantity is similar, and when a value is large, it is shown that it is a dissimilarity. Here, a function which calculates the dissimilarity nature of two segment S_1 about the characteristic quantity F and S_2 is defined as dissimilarity nature metrics d_F (S_1, S_2). Such a function satisfies a relation given by the following formulas (2).

[0071]

[Equation 2]

$$d_F(S_1, S_2) = 0 \qquad (S_1 = S_2 \text{のとき})$$
$$d_F(S_1, S_2) \geq 0 \qquad (\text{全ての} S_1, S_2 \text{について}) \qquad \cdots (2)$$
$$d_F(S_1, S_2) = d_F(S_2, S_1) \qquad (\text{全ての} S_1, S_2 \text{について})$$

[0072]By the way, although some of dissimilarity nature metrics are applicable only to a certain specific characteristic quantity, "G. Ahanger and T.D.C. Little, A survey of technologies forparsing and indexing digital video, J. of Visual Communication and Image Representation. 7: 28-4, 1996", and "L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons,. Generally many dissimilarity nature metrics are applicable to measuring the similarity about the characteristic quantity expressed as a point in an n-space as indicated to 1990." The example is Euclidean distance, an inner product, L1 distance, etc. Here, since especially L1 distance acts effectively to various characteristic quantity containing characteristic quantity, such as a histogram and imagery correlation, the video voice processing unit 10 introduces L1 distance. Here, when two n vectors are set to A and B, L1 distance d_{L1} (A, B) between A and B is given with a following formula (3).

[0073]

[Equation 3]

$$d_{L1}(A,B) = \sum_{i=1}^{n} |A_i - B_i| \qquad \cdots (3)$$

[0074]Here, the character i with the bottom shows each i-th element of n vectorA and B.

[0075]The video voice processing unit 10 extracts the static characteristic quantity in various times in a segment as characteristic quantity showing dynamic features, as mentioned above. And in order to determine the similarity between the two extracted amounts of dynamic features, the dissimilarity nature metrics between the amounts of static features used as the foundation are used for the video voice processing unit 10 as the dissimilarity nature metrics. It is best that the dissimilarity nature metrics of these amounts of dynamic features are determined using a pair of dissimilarity nature value of the most similar amount of static features chosen from each amount of dynamic features in many cases. In this case, the dissimilarity nature metrics between two extracted amount SF_1 of dynamic features and SF_2 are defined like a following formula (4).

[0076]

[Equation 4]

$$d_s(SF_1, SF_2) = \min_{F_1 \in SF_1, F_2 \in SF_2} d_F(F_1, F_2) \qquad \cdots (4)$$

[0077]Here, function $d_F(F_1, F_2)$ in an upper type (4) shows dissimilarity nature metrics about the amount F of static features used as the foundation. It is very good in the maximum or average value instead of taking the minimum of the dissimilarity nature of characteristic quantity depending on the case.

[0078]By the way, when the video voice processing unit 10 determines the similarity of a segment, just its single characteristic quantity is insufficient, and it needs to combine information from characteristic quantity of a large number about the same segment in many cases. As this one method, the video voice processing unit 10 calculates dissimilarity nature based on various characteristic quantity as a combination with dignity of each characteristic quantity. That is, when k characteristic quantity F_1, F_2, ..., F_k exist, dissimilarity nature metrics $d_F(S_1, S_2)$ about combined characteristic quantity which is expressed with a following formula (5) is used for the video voice processing unit 10.

[0079]

[Equation 5]

$$d_F(S_1, S_2) = \sum_{i=1}^{k} w_i d_{Fi}(S_1, S_2) \quad \cdots (5)$$

[0080]Here, $\{w_i\}$ is a weighting factor used as $\mathrm{sigma}_i w_i = 1$.

[0081]As mentioned above, the video voice processing unit 10 can calculate dissimilarity nature metrics using characteristic quantity extracted in the step S2 in drawing 5, and can measure similarity between the segments concerned.

[0082]Below, chain detection in step S4 in drawing 5 is explained. The video voice processing unit 10 detects a similar chain showing relation between similar segments using dissimilarity nature metrics and extracted characteristic quantity. Here, first, here a similar chain of some types is defined and an algorithm for detecting a similar chain of each type is explained concretely.

[0083]by the way, a type of a similar chain defined below is mutually-independent respectively -- since the bottom is a thing, in the video voice processing unit 10, one chain is able to belong to two or more types. Here, such a chain will be called combining a defined type name. For example, a partial uniform link chain is local, is uniform, and shows a thing of a link similar chain so that it may mention later.

[0084]Now, a type of a similar chain is divided roughly into what has restrictions in a relation between similar segments which the similar chain concerned contains, and a thing which has restrictions in structure of the similar chain concerned. Suppose that the chain C expresses a series of segment Si_1, ..., Si_m in the following definitions. Index i_k expresses a segment number in the original video data of the segment here, and subscript k to i means that the segment is located on a time-axis in the similar chain concerned the k-th. On a time-axis, these segments of a series of shall always be set

in order, and are $i_k<i_{k+1}$ about all the k= 1, ..., m−1. |C| shall express the length of a chain and C^{start} and C^{end} shall express start time and finish time of the chain C in a video data, respectively. More correctly, start time of the chain C is the start time of the first segment in the chain C, and finish time of the chain C is the finish time of a segment of the last in the chain C. a case where a certain segment is set to A further again -- the similar segment -- A' and A -- it expresses with '', A''', and ... Finally it supposes that it is in a state where those dissimilarity nature metrics are smaller than a dissimilarity nature threshold mentioned later that two segments are similar, and this is expressed with similar (S_1, S_2).

[0085]As a similar chain which has restrictions, a basic similar chain, a link similar chain, and a periodic chain have a relation between similar segments which the similar chain concerned contains.

[0086]First, although it is a basic similar chain, this is the chain C with which all the segments were mutually similar, as shown in drawing 7. There are no structural restrictions in a basic similar chain. In many cases, this basic similar chain is obtained as a result of a grouping algorithm for carrying out grouping of the similar segment, or a clustering algorithm.

[0087]On the other hand, as it is indicated in drawing 8 as a link similar chain, an adjoining segment is the mutually similar chain C. That is, in a link similar chain, it is similar (S_k, S_{k+1}) about all the k= 1, ---, and |C|−1. A' from a definition of a similar segment which this link similar chain mentioned above, and A -- it can be described as '', A''', and ...

[0088]As it is indicated in drawing 9 as a periodic chain, each segment is a segment of the m−th back, and similar chain C_{cyclic}. That is, in a periodic chain, it is similar (S_k, S_{k+1}) about all the k= 1, ..., $|C_{cyclic}|$−1. If it puts in another way, a periodic chain is constituted as an approximate repetition of a series of m segments. From this, a periodic chain, S -- $_1$ -- S -- $_2$ - - - S -- $_m$ -- S -- $_1$ -- ' -- S -- $_2$ -- ' ... S -- $_m$ -- ' -- S -- $_1$ -- ' -- ' -- S -- $_2$ -- ' -- ' ... S -- $_m$ -- ' -- ' -- ** -- S -- $_1$ -- ' -- ' -- ' -- S -- $_2$ -- ' -- ' -- ' .... It can be described as S_m'''.

[0089]On the other hand, there are a partial chain and a uniform chain as a similar chain which has structural restrictions.

[0090]Here, in each set of an adjoining segment, a partial chain is the chain C whose time interval between segments is smaller than predetermined time, as mentioned above. Namely, if the maximum of a time interval permitted between two segments in a chain is expressed in a partial chain as gap, It is $i_{k+1}-i_k<=$gap to segment Si_k and Si_{k+1} which adjoin about all the k= 1, ..., |C|−1.

[0091]When a segment in a chain appears in an almost equal time interval, this can serve as a leading index of important video structure, but such a chain C is defined as a uniform chain. Here, homogeneous uniformity (C) of the chain C is defined like as a thing which is shown in a following formula (6) and which standardized average value of

a gap of a time interval from regular–intervals time by the length of the chain.

[0092]

[Equation 6]

$$uniformity(C) = \frac{\sum_{i=1}^{|c|-1} \left| \left(S_{i+1}^{start} - S_{i}^{start}\right) - \frac{\left(C^{end} - C^{start}\right)}{|C|} \right|}{|C| \cdot \left|C^{end} - C^{start}\right|} \quad \cdots (6)$$

[0093]Homogeneous uniformity (C) of the chain C shown by an upper formula (6) takes the value of the range of 0 to 1, and when the value is small, it shows that it is close to distribution with uniform time interval distribution of a segment. When the value of this homogeneous uniformity (C) is smaller than a homogeneous predetermined threshold, it is considered that the chain C is a uniform chain.

[0094]Hereafter, in the video voice processing unit 10, the processing for detecting each of such various chains is explained.

[0095]In order to detect a basic similar chain mentioned above, batch clustering technology or serial clustering technology is used for the video voice processing unit 10.

[0096]Batch clustering technology is the art of detecting a chain collectively. However, in order to apply this art, before performing chain detection, it is necessary to end all the video division. One serial clustering technology is the art of detecting a chain sequentially, and supposing video division and characteristic quantity extraction are performed sequentially again, it will become possible to conduct video analysis sequentially, playing a video data. there is sufficient count ability for the video voice processing unit 10 -- if it becomes -- this successive chain detection -- real time -- if it puts in another way, a chain is detectable while taking in or recording a video data. However, successive video analysis may produce a problem for the accuracy. That is, in the case of a successive method, there is no global information for determining the optimal chain structure, and since it is still more sensitive to an entry sequenced foreword of a segment, a result of low quality may be produced.

[0097]When using batch clustering technology, the video voice processing unit 10 detects a basic similar chain by passing through two processes, as shown in drawing 10.

[0098]First, the video voice processing unit 10 detects a candidate chain in Step S11. Namely, the video voice processing unit 10 detects a similar segment in a video data, and summarizes it as a cluster. A cluster group of a segment obtained by this serves as an initial candidate when detecting a basic similar chain.

[0099]When the video voice processing unit 10 asks for an initial candidate of a similar

chain, arbitrary clustering technologies can be used for it, but. Here, To "L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990." A hierarchical clustering method (hierarchical clustering method) indicated will be used. This algorithm starts by summarizing two most similar segments as one pair first, and summarizes one pair of a cluster which was most similar in each stage after another using similarity metrics between clusters. In this algorithm, dissimilarity nature metrics d_c (C_1, C_2) between two cluster C_1 and C_2 is defined as minimum dissimilarity nature between two segments contained in each cluster, as shown in a following formula (7).

[0100]

[Equation 7]

$$d_c\left(C_1, C_2\right) = \min_{s_1 \in c_1, s_2 \in c_2} d_s\left(S_1, S_2\right) \quad \cdots (7)$$

[0101]In the video voice processing unit 10, the maximum function or an average function may be used instead of the minimum function shown by an upper formula (7) if needed.

[0102]By the way, this hierarchical clustering method will summarize all the segments contained in a video data in a single group, when there are no restrictions temporarily. Then, as shown in drawing 11, the video voice processing unit 10 introduces dissimilarity nature threshold $delta_{sim}$, and judges whether a certain segment is similar to the segment of another side by comparison with this dissimilarity nature threshold $delta_{sim}$. As it is indicated in the figure as dissimilarity nature threshold $delta_{sim}$ here, when how many two segments are similar, it is a threshold which determines whether to regard it as the thing belonging to the same chain. And the video voice processing unit 10 summarizes the segment to the cluster in the range in which the dissimilarity nature of all the cluster pair does not exceed this dissimilarity nature threshold $delta_{sim}$.

[0103]It may be made for the video voice processing unit 10 to set up dissimilarity nature threshold $delta_{sim}$ by a user, and it may determine it automatically. However, when using a fixed value as dissimilarity nature threshold $delta_{sim}$, it will depend for the optimum value on the contents of the video data. For example, in the case of a video data which has the image contents which were varied, dissimilarity nature threshold $delta_{sim}$ needs to be set as a high value. In the case of a video data which, on the other hand, has image contents with little change, dissimilarity nature threshold $delta_{sim}$ needs to be set as a low value. Generally a cluster number detected when dissimilarity nature threshold $delta_{sim}$ is high decreases here, and when dissimilarity nature threshold $delta_{sim}$ is low, a cluster number detected has the character to increase.

[0104]From this, in the video voice processing unit 10, when determining suitable

dissimilarity nature threshold $delta_{sim}$ influences the performance, it becomes important. Therefore, in the video voice processing unit 10, to set up dissimilarity nature threshold $delta_{sim}$ by a user, after taking having mentioned above into consideration, it is necessary to set up. On the other hand, the video voice processing unit 10 can also determine effective dissimilarity nature threshold $delta_{sim}$ automatically by a method shown below.

[0105]For example, the video voice processing unit 10 can obtain dissimilarity nature threshold $delta_{sim}$ as the one method using a statistics value called average value and a median (median) in distribution of dissimilarity nature between (n) (n−1) / 2 segment pairs. When average value and standard deviation of dissimilarity nature in all the segment pairs are set to mu and sigma now, respectively, dissimilarity nature threshold $delta_{sim}$ can be expressed with form of amu+bsigma. Here, a and b are constants and it has found out giving a result with respectively good setting it as 0.5 and 0.1.

[0106]On practical use, the video voice processing unit 10, What is necessary is not to ask for dissimilarity nature between them, and for the average value mu and the standard deviation sigma to choose from all the segment pair sets at random a segment pair which is sufficient for giving a result sufficiently near a true value, and just to ask for the dissimilarity nature about all the segment pairs. The video voice processing unit 10 can obtain suitable dissimilarity nature threshold $delta_{sim}$ automatically by using the average value mu acquired by doing in this way, and the standard deviation sigma. That is, the video voice processing unit 10 can determine suitable dissimilarity nature threshold $delta_{sim}$ automatically by extracting a number of dissimilarity nature of a segment pair given by Cn, when total of a segment pair is set to n and arbitrary small constants are set to C for example.

[0107]As the video voice processing unit 10 had been shown so far, after clustering a segment, it can obtain an initial candidate of a basic similar chain by rearranging a segment contained in each cluster concerned in each cluster.

[0108]By the way, video structure with the actual many of a chain candidate who detected in the step S11 in drawing 10 is unrelated. From this, the video voice processing unit 10 needs to determine whether to be an important chain with which which chain candidate makes a skeleton of video structure, or be a chain relevant to video structure. Therefore, the video voice processing unit 10 performs chain filtering using quality metrics corresponding to a numerical standard which shows quality of a chain in Step S12. Namely, the video voice processing unit 10 measures a chain candidate's importance and relevance in video structure analysis, and outputs them as a result of chain detection of only a chain candidate who exceeds a predetermined quality metrics threshold. Here, although the simplest example as a relevance measurement function used by filtering is a Boolean function which shows whether a chain candidate is accepted, as for the video voice processing unit 10, a more

complicated relevance measurement function may be used if needed.

[0109]By the way, in the video voice processing unit 10, chain length, chain density, chain strength, etc. are used as chain quality metrics.

[0110]First, although it is chain length, this is defined as the number of segments which one chain holds. Here, generally it is when chain length is small that the video voice processing unit 10 can use this chain length as those chain quality metrics, and it depends on usually regarding as a noise being possible. For example, when a certain chain has only unisegment, it does not have any information. That is, in quality metrics based on chain length, the minimum of the number of segments which a chain should hold will be given as the restrictions.

[0111]Next, although it is chain density, this is defined as a ratio of the total number of segments which a certain chain holds, and the total number of segments in subregion of a video data which the chain occupies. This depends on that it may be more desirable to exist intensively in a segment of time to which a chain was restricted. In this case, this chain density should just be used for the video voice processing unit 10 as those chain quality metrics.

[0112]Finally, although it is chain strength, this is an index which shows whether each segment in a chain is how much similar mutually, and it considers that the chain has high intensity, so that the segment concerned is mutually similar. In the video voice processing unit 10, about a method of measuring this chain strength. A large number exist including a similarity measuring method in a chain shown below, a method of taking average value of dissimilarity nature between all the possible segment pairs, or a method of taking the maximum of dissimilarity nature between all the possible segment pairs.

[0113]As an example, the video voice processing unit 10 shows a case where chain strength is measured with a similarity measuring method in a chain. Here, a similarity measuring method in a chain is a method of expressing the similarity of a segment which constitutes a chain as average value of the dissimilarity nature of each segment and the most typical segment that the chain contains. As an example of a typical segment, a center-of-gravity (centroid) segment of a chain is mentioned. If a center-of-gravity segment in the chain C is made into $S_{centroid}$ now, this center-of-gravity segment $S_{centroid}$ will be defined by following formula (8).

[0114]

[Equation 8]

$$S_{centroid} = \underset{S_A \in C}{argmin} \frac{1}{|C|} \sum_{S_B \in C} d_F(S_A, S_B) \quad \cdots (8)$$

[0115]Here, argmin in an upper type (8) means choosing input S_A**C which makes the value of the formula of an evaluation object the minimum.

[0116]From this, when chain strength is made into $d_{centroid}$, this chain strength $d_{centroid}$ is expressed like a following formula (9).

[0117]

[Equation 9]

$$d_{centroid} = \frac{1}{|C|} \sum_{S \in C} d_F \left(S , S_{centroid} \right) \quad \cdots (9)$$

[0118]Now, the video voice processing unit 10 performs chain filtering using the chain quality metrics mentioned above by a series of processings as concretely shown in drawing 12.

[0119]First, in Step S21, the video voice processing unit 10 makes filtering chained list $C_{filtered}$ a nil state while initializing chained list C_{list} with a candidate chain.

[0120]Then, the video voice processing unit 10 distinguishes whether chained list C_{list} is a nil state in Step S22.

[0121]Here, when chained list C_{list} is a nil state, the video voice processing unit 10 ends a series of processings from the target candidate chain not existing.

[0122]On the other hand, when chained list C_{list} is not a nil state, in Step S23, the video voice processing unit 10 makes a certain chain C the element of the beginning of chained list C_{list}, and removes the chain C from chained list C_{list}.

[0123]Then, the video voice processing unit 10 calculates chain quality metrics about the chain C in Step S24.

[0124]And the video voice processing unit 10 distinguishes whether these chain quality metrics are larger than a quality metrics threshold in Step S25.

[0125]Here, when chain quality metrics are smaller than a quality metrics threshold, the video voice processing unit 10 shifts processing to Step S22, and processing about another chain is performed again.

[0126]On the other hand, when chain quality metrics are larger than a quality metrics threshold, the video voice processing unit 10 adds the chain C to filtering chained list $C_{filtered}$ in Step S26.

[0127]And the video voice processing unit 10 distinguishes whether chained list C_{list} is a nil state in Step S27.

[0128]Here, when chained list C_{list} is a nil state, the video voice processing unit 10 ends a series of processings from the target candidate chain not existing.

[0129]On the other hand, when chained list C_{list} is not a nil state, the video voice processing unit 10 shifts processing to Step S23. Thus, the video voice processing unit 10 repeats processing until chained list C_{list} will be in a nil state.

[0130]By such a series of processings, the video voice processing unit 10 can perform chain filtering, and can determine which chain is a chain relevant to whether it is an important chain which makes a skeleton of video structure, and video structure.

[0131]As mentioned above, the video voice processing unit 10 can detect a basic

similar chain using such batch clustering technology.

[0132]By the way, the video voice processing unit 10 can also detect a basic similar chain as an option using serial clustering technology mentioned above with batch clustering technology. That is, the video voice processing unit 10 processes every one segment in a video data according to order of the input, and repeats and updates a chain candidate list. Also in this case, like batch clustering technology, the video voice processing unit 10 divides a main process of chain detection into two steps, and is performed. That is, the video voice processing unit 10 detects a cluster of a similar segment first using a clustering algorithm one by one. Next, the video voice processing unit 10 filters a detected cluster using the same chain quality metrics as batch clustering technology. Here, in a point advanced in a stage where filtering of a chain is early, as filtering processing at the time of using clustering technology one by one, the video voice processing unit 10 differs from a case of batch clustering technology.

[0133]Now, in clustering technology, when clustering a segment, a clustering algorithm is used one by one. By the way, generally almost all serial clustering is performed to partial optimum. That is, with a clustering algorithm, it is judged locally whether the segment is assigned to the existing cluster whenever a new segment is inputted, or the new cluster containing only the segment is generated one by one. There are some which update the cluster division itself whenever a new segment is inputted, in order to prevent a bias effect accompanying an entry sequenced foreword of a segment as an on the other hand more elaborate serial clustering algorithm. About such an algorithm. "J. Roure and L. Talavera, Robust incremental clustering with bad instance orderings:a new strategy,In Proceedings of the Sixth Iberoamerican. Conference on Artifical. Intelligence, IBERAMIA-98. Pages 136-147. Lisbon, Portugal. Helder Coelho ed.,LNAI vol. 1484. Springer Verlag, 1998." ****** can be referred to.

[0134]The video voice processing unit 10 performs processing as shown in drawing 13 as an example of a clustering algorithm one by one. Here, a video data divided into a segment considers it as segment S_1, ---, and a thing that has S_n. Here, a series of processings also including a process of chain analysis are explained.

[0135]First, as shown in the figure, in Step S31, the video voice processing unit 10 initializes chained list C_{list} to a nil state, and sets segment number i as 1 in Step S32.

[0136]Next, the video voice processing unit 10 distinguishes whether segment number i is smaller than the total segment n [several] in Step S33.

[0137]Here, since the target segment [i / segment number / processing unit / 10 / video voice] when larger than the total segment n [several] does not exist, a series of processings are ended.

[0138]On the other hand, segment number i in being smaller than the total segment n [several], In Step S34, the video voice processing unit 10 incorporates segment S_1, and distinguishes whether chained list C_{list} is a nil state in Step S35 at segment S_i, i.e., here.

[0139]Here, when chained list C_{list} is a nil state, the video voice processing unit 10 shifts processing to Step S42.

[0140]On the other hand, when chained list C_{list} is not a nil state, the video voice processing unit 10 calculates chain C_{min} whose dissimilarity nature to segment S_1 is the minimum in Step S36. Here, chain C_{min} is defined like a following formula (10).

[0141]

[Equation 10]

$$C_{min} = \underset{C \in C_{list}}{argmin}\, d_{SC}(C, S_i) \qquad \cdots (10)$$

[0142]In an upper type (10), d_{SC} (C, S) expresses the dissimilarity nature metrics between the chain C and the segment S, and is given with a following formula (11).

[0143]

[Equation 11]

$$d_{SC}(C, S) = \underset{s_1 \in C}{min}\, d_F(S, S_i) \qquad \cdots (11)$$

[0144]In the upper type (7) which is the similarity metrics defined in batch clustering technology, this is equivalent to what made the 2nd argument the cluster having contained only the segment concerned. Below, suppose that minimum dissimilarity nature d_{SC} (C_{min}, S_i) between chain C_{min} and segment S_i is only expressed as d_{min}.

[0145]Next, the video voice processing unit 10 distinguishes whether minimum dissimilarity nature d_{min} is smaller than dissimilarity nature threshold $delta_{sim}$ in Step S37 using dissimilarity nature threshold $delta_{sim}$ which was explained in the case of batch clustering technology.

[0146]Here when minimum dissimilarity nature d_{min} is larger than dissimilarity nature threshold $delta_{sim}$, In [the video voice processing unit 10 shifts to processing of Step S42, generate new chain C_{new} which has only the segment S_i concerned as an only element segment, and] Step S43, New chain C_{new} is added to chained list C_{list}, and it shifts to processing of Step S39.

[0147]On the other hand, when minimum dissimilarity nature d_{min} is smaller than dissimilarity nature threshold $delta_{sim}$, in Step S38, as for the video voice processing unit 10, the segment S_i concerned is added to chain C_{min}. That is, the video voice processing unit 10 is made into C_{min}<-C_{min}**S_i.

[0148]And the video voice processing unit 10 filters a chain in Step S39. That is, as mentioned above, about each element chain C**C_{list}, the video voice processing unit 10 measures quality of the chain C, chooses only a chain which has quality metrics which exceed a quality metrics threshold, and adds this to chained list $C_{filtered}$.

[0149]The video voice processing unit 10 analyzes a chain sequentially in Step S40.

That is, the video voice processing unit 10 lets filtered chained list $C_{filtered}$ in the time pass to an analysis module.

[0150]And in Step S41, the video voice processing unit 10 adds 1 to segment number i, and shifts to processing of Step S33.

[0151]Thus, the video voice processing unit 10 until segment number i becomes larger than the total segment n [several], A series of above processings are repeated and each element chain of chained list C_{list} at the time of segment number i becoming larger than the total segment n [several] is detected as a basic similar chain.

[0152]A series of processings shown in the figure are premised on the total segment n [several] contained in an inputted video data being known. However, generally the total segment n [several] is not given beforehand in many cases. In that case, the clustering algorithm should just distinguish continuation or an end of processing by whether there is any input of a segment succeedingly in Step S33 in the said figure one by one.

[0153]By such a series of processings, the video voice processing unit 10 can detect a basic similar chain which used clustering technology one by one.

[0154]Processing which detects next a link similar chain mentioned above is explained. Detection of a link similar chain in the video voice processing unit 10 can be considered as a special case of basic similar chain detection. The video voice processing unit 10 performs processing as shown in drawing 14 as a link similar chain detecting method which used a clustering algorithm one by one. Here, a video data divided into a segment assumes that it has segment S_1, ..., S_n. Here, a series of processings also including a process of chain analysis are explained.

[0155]As shown in the figure, in Step S51, the video voice processing unit 10 initializes chained list C_{list} to a nil state, and sets segment number i as 1 in Step S52.

[0156]Next, the video voice processing unit 10 distinguishes whether segment number i is smaller than the total segment n [several] in Step S53.

[0157]Here, since the target segment [i / segment number / processing unit / 10 / video voice] when larger than the total segment n [several] does not exist, a series of processings are ended.

[0158]On the other hand, segment number i in being smaller than the total segment n [several], In Step S54, the video voice processing unit 10 incorporates segment S_1, and calculates chain C_{min} whose dissimilarity nature to segment S_1 is the minimum in Step S55 at segment S_i, i.e., here. Here, chain C_{min} is defined like a following formula (12).

[0159]

[Equation 12]

$$C_{\min} = \underset{c \in C_{list}}{argmin}\, d_{SC}(C, S_i) \qquad \cdots (12)$$

[0160]In an upper type (12), although d_{SC} (C, S) expresses the dissimilarity nature metrics between the chain C and the segment S too, in link similar chain detection, this dissimilarity nature metrics d_{SC} (C, S) is given with a following formula (13).

[0161]

[Equation 13]

$$d_{SC} = (C, S)\, d_F\left(S_{|C|}, S_i\right) \quad \cdots (13)$$

[0162]That is, unlike an upper type (11) which was used on the occasion of detection of a basic similar chain and which is dissimilarity nature metrics, dissimilarity nature metrics d_{SC} (C, S) is given as dissimilarity nature between the segment concerned and an element segment of the last in the chain C.

[0163]Next, the video voice processing unit 10 distinguishes whether minimum dissimilarity nature d_{min} is smaller than dissimilarity nature threshold $delta_{sim}$ in Step S56 using dissimilarity nature threshold $delta_{sim}$ which was mentioned above.

[0164]Here when minimum dissimilarity nature d_{min} is larger than dissimilarity nature threshold $delta_{sim}$, In [the video voice processing unit 10 shifts to processing of Step S61, generate new chain C_{new} which has only the segment S_i concerned as an only element segment, and] Step S62, New chain C_{new} is added to chained list C_{list}, and it shifts to processing of Step S58.

[0165]On the other hand, when minimum dissimilarity nature d_{min} is smaller than dissimilarity nature threshold $delta_{sim}$, in Step S57, as for the video voice processing unit 10, the segment S_i concerned is added to an end of chain C_{min}. That is, the video voice processing unit 10 is made into C_{min}<-C_{min} and S_i.

[0166]And the video voice processing unit 10 filters a chain in Step S58. That is, as mentioned above, about each element chain C**C_{list}, the video voice processing unit 10 measures quality of the chain C, chooses only a chain which has quality metrics which exceed a quality metrics threshold, and adds this to chained list $C_{filtered}$. The video voice processing unit 10 can also skip this process.

[0167]The video voice processing unit 10 analyzes a chain sequentially in Step S59. That is, the video voice processing unit 10 lets filtered chained list $C_{filtered}$ in the time pass to an analysis module.

[0168]And in Step S60, the video voice processing unit 10 adds 1 to segment number i, and shifts to processing of Step S53.

[0169]Thus, the video voice processing unit 10 until segment number i becomes larger than the total segment n [several], A series of above processings are repeated and each element chain of chained list C_{list} at the time of segment number i becoming larger than the total segment n [several] is detected as a link similar chain.

[0170]By such a series of processings, the video voice processing unit 10 can detect a link similar chain using such serial clustering technology.

[0171]A series of processings shown in the figure are premised on the total segment n [several] contained in an inputted video data being known. However, generally the total segment n [several] is not given beforehand in many cases. In that case, the clustering algorithm should just distinguish continuation or an end of processing by whether there is any input of a segment succeedingly in Step S53 in the said figure one by one.

[0172]Processing which detects next a periodic chain mentioned above is explained. It can be considered that periodic chain C_{cyclic} is that $\{C_1, ..., C_k\}$ whose k different basic similar chains or link similar chains settled. Hereafter, C (S_i) presupposes a segment in periodic chain C_{cyclic} S_1, ---, and that it is described as S_n and the chain number 1 of appearance origin of segment S_i, ..., k are shown. from this, C_{cyclic} is a periodic chain -- if it becomes -- C (S_1), C (S_2), and --- C (S_n) -- a row of a series of chain numbers, It will be described in i_l, ---, i_k, i_1, ..., i_k, ..., i_l, ---, and form of i_k. Here, i_l, ..., i_k are permutation of the chain number 1, ..., k, and arbitrary rows which will not overlap if it puts in another way by the one cycle. Below, the number of segments contained in 1 cycle decides periodic chain i_l which is one, i_l, ---, and to call i_l a fundamental-period chain.

[0173]By the way, since cyclic structures in a video data are not the thoroughly congruous things and each cycle is usually approximate, the video voice processing unit 10 looks for an approximate periodic chain in a video data by a series of processings as shown in drawing 15. Here, constraints that the video voice processing unit 10 must have a uniform fundamental-period chain which becomes origin of it if needed can be added. Here, processing performed on a basis of these constraints is explained.

[0174]First, in [as the video voice processing unit 10 is shown in the figure] Step S71 and Step S72, A fundamental-period chain contained in a video data is detected, an initial chained list is generated based on it, and an initial chained list is updated so that all the fundamental-period chains further contained in an initial chained list may fulfill constraints of a uniform chain.

[0175]That is, the video voice processing unit 10 calculates initial chained list C_{list} in Step S71 using an algorithm which detects a basic similar chain or a link similar chain mentioned above.

[0176]And in Step S72, about each chain C contained in an initial chained list, the video voice processing unit 10 checks that homogeneity, and the chain C divides it into two or more uniform subchains with which that time interval serves as the maximum in this chain C, when not uniform. Then, the video voice processing unit 10 is filtered using chain quality metrics which were explained in an algorithm which detects a basic similar chain or a link similar chain which mentioned an obtained uniform subchain above, A selected uniform subchain is added to initial chained list C_{list}.

[0177]Next, in Step S73, the video voice processing unit 10 out of chained list C_{list}. It

overlaps in time and one pair of crossing chains, i.e., **C_1, chain C_1 C_2|[C_1^{start}, C_1^{end}] ** [C_2^{start}, C_2^{end}] Becoming, and C_2 are calculated.

[0178]And the video voice processing unit 10 distinguishes whether such duplicate chain C_1 and C_2 exist in Step S74.

[0179]Here, when duplicate chain C_1 and C_2 do not exist, the video voice processing unit 10 ends a series of processings as that in which chained list C_{list} has already contained two or more periodic chains.

[0180]On the other hand, when duplicate chain C_1 and C_2 exist, In order to determine whether the video voice processing unit 10 constitutes one periodic chain two chain C_1 and whose C_2 settled in Step S75 thru/or Step S78, Compatibility between each cycle is evaluated in a periodic chain which doubled the two periodic chains.

[0181]That is, in Step S75, the video voice processing unit 10 doubles two chain C_1 and C_2, and forms new periodic chain C_M. Here, suppose that a segment in chain C_M is expressed S_1, S_2, ..., $S_{|CM|}$.

[0182]Then, the video voice processing unit 10 sets the chain number C of appearance origin of segment S_1 (S_1) to C in Step S76, In row [of a chain number] C (S_1), C (S_2), ..., C ($S_{|CM|}$), for every generating of C. That is, chain C_M is decomposed into subchain C_M^1, C_M^2, ..., C_M^k bordering on just before a segment belonging to the same chain as segment S_1 appears. As a result, the video voice processing unit 10 obtains a list of subchains as shown in a following formula (14).

[0183]

[Equation 14]

$$\begin{aligned} C_M^1 &= S_1, \cdots, S_i, \\ C_M^2 &= S_{i_1+1}, \cdots, S_{i_2}, \\ &\vdots \\ C_M^k &= S_{i_{k-1}+1}, \cdots, S_{i_k}, \end{aligned} \qquad \cdots (14)$$

[0184]In an upper type (14), $C(S_{ij+1}) = C(S_1)$ is realized about all the C_M^j so that clearly from this operation.

[0185]Then, the video voice processing unit 10 finds subchain C_M^{cycle} with the highest frequency of occurrence in Step S77. That is, the video voice processing unit 10 performs processing as shown in a following formula (15).

[0186]

[Equation 15]

$$C_M^{cycle} = \underset{C_M^k}{argmax} \left| \left\{ C_M^i \,\middle|\, C_M^i = C_M^k, i \in \{1, \cdots, k\} \right\} \right| \qquad \cdots (15)$$

[0187]And the video voice processing unit 10 evaluates whether subchain C_M^{cycle} with the highest frequency of occurrence can become one cycle of the original chain C_M in Step S78. Namely, as shown in a following formula (16), the video voice processing unit 10 the consistency coefficient mesh, It defines by the ratio to the subchain total of the frequency of occurrence of C_M^{cycle} calculated at Step S76, and it is distinguished whether this consistency coefficient exceeds a predetermined threshold in continuing Step S79.

[0188]

[Equation 16]

$$mesh = \frac{\left| \left\{ C_M^i \,\middle|\, C_M^i = C_M^{cycle}, i \in \{1, \cdots, k\} \right\} \right|}{k} \quad \cdots (16)$$

[0189]Here, when the consistency coefficient is not over the threshold, the video voice processing unit 10 shifts to processing of Step S73, and repeats the same processing in quest of other duplicate chains.

[0190]On the other hand, when the consistency coefficient is over the threshold, In [the video voice processing unit 10 removes chain C_1 and C_2 from chained list C_{list} in Step S80, and] Step S81, Chain C_M is added to chained list C_{list}, and it shifts to processing of Step S73.

[0191]The video voice processing unit 10 by repeating such a series of processings until the chain which overlaps about all the periodic chains contained in chained list C_{list} stops existing, Chained list C_{list} containing a final periodic chain can be obtained.

[0192]As mentioned above, the video voice processing unit 10 can detect the various chains of a similar segment using dissimilarity nature metrics and the extracted characteristic quantity.

[0193]Below, chain analysis in the step S5 in drawing 5 is explained. The video voice processing unit 10 determines and outputs local video structure and/or global video structure of a video data using a detected chain. Here, although fundamental structural patterns by which it is generated in a video data are detected, a concrete example is given and explained about using a result of chain analysis how.

[0194]First, a scene which is local structural patterns by which it is generated in a video data is explained.

[0195]As mentioned above, a scene is a unit of the most fundamental local video structure positioned by higher rank from a level of a segment, and comprises a series of semantically related segments. The video voice processing unit 10 can detect these scenes using a chain. In scene detection in the video voice processing unit 10,

conditions which a chain should fulfill are that a time interval between segments which continued mutually exceeds a certain defined value which is called a time threshold about no segments which the chain contains. Here, a chain which fulfills this condition is called a partial chain.

[0196]The video voice processing unit 10 performs a series of processings as shown in drawing 16, in order to detect a scene using a chain.

[0197]First, the video voice processing unit 10 asks for a partial chained list in Step S91 thru/or Step S94, as shown in the figure.

[0198]That is, the video voice processing unit 10 asks for 1 set of initial chained lists in Step S91 using basic similar chain detection algorithms mentioned above.

[0199]Next, in Step S92, about each chain C in an initial chained list for which it asked, when the chain C is not a partial chain, the video voice processing unit 10. The chain C is decomposed into a row of partial subchain $C=C_1$ and ... which are the longest, and C_n in the condition range of a partial chain.

[0200]Then, the video voice processing unit 10 removes the chain C from a chained list in Step S93.

[0201]The video voice processing unit 10 adds each subchain C_i to a chained list in Step S94. After this process is completed, all the chains become local.

[0202]Next, in Step S95, the video voice processing unit 10 out of a chained list. One pair of duplicate chain C_1, C_2 which cross in time, That is, $**C_1$, chain C_1 which is $C_2|[C_1^{start}, C_1^{end}]$ ** $[C_2^{start}, C_2^{end}]$, and C_2 are calculated.

[0203]Then, the video voice processing unit 10 distinguishes whether such duplicate chain C_1 and C_2 exist in Step S96.

[0204]Here, when duplicate chain C_1 and C_2 do not exist, the video voice processing unit 10 ends a series of processings as that in which one scene exists for every chain contained in a chained list.

[0205]On the other hand, when duplicate chain C_1 and C_2 exist, in Step S97, the video voice processing unit 10 doubles duplicate chain C_1 and C_2, and forms new chain C_M.

[0206]In Step S98, the video voice processing unit 10 removes chain C_1 and C_2 duplicate from a chained list, adds chain C_M, shifts to processing of Step S95 again after that, and repeats the same processing.

[0207]When a duplicate chain stops existing in a chained list as a result of doing in this way, one scene will exist for each [which was contained in a chained list obtained eventually] chain of every. A boundary of scene S_j corresponding to chain C_j is given by C^{start} and C^{end}.

[0208]By the way, although some segments remain without being assigned to any chains, the video voice processing unit 10 summarizes such a segment that remained between two detected scenes as a default, and makes it one scene.

[0209]By such a series of processings, the video voice processing unit 10 can detect a scene which is the local structural patterns in a video data by using a chain.

[0210]A case where such processing is applied to a conversation scene previously shown in drawing 2 is considered. In this case, the video voice processing unit 10 asks for a partial chain about each of a speaker's segment in Step S91 thru/or Step S94. And in Step S97, the video voice processing unit 10 will pack these chains, and will form a single large chain showing the whole scene.

[0211]Thus, the video voice processing unit 10 can detect a scene in a conversation scene.

[0212]In the video voice processing unit 10, when a scene is detected, cautions are taken not to contain all segments in a scene in a chain.

[0213]The video voice processing unit 10 can also detect a scene sequentially by performing sequentially an algorithm mentioned above.

[0214]Below, a case where a news item is detected is explained as global structural patterns.

[0215]As mentioned above, the news item starts with an introductory sentence by an anchor first, for example, and a news program has the cyclic structures that one or more reports from the spot continue. That is, it can be considered that such video structure is the simple cyclic structures which made one cycle just before [from an anchor shot to] the next anchor shot.

[0216]The video voice processing unit 10 performs a series of processings in which an outline is shown in drawing 17, in order to detect a news item automatically using a chain.

[0217]First, the video voice processing unit 10 detects a periodic chain in Step S101 using periodic chain detection algorithms mentioned above, as shown in the figure. By performing this process, the video voice processing unit 10 can obtain a list of periodic chains. Here, each cycle may express a news item and does not need to express it.

[0218]Next, the video voice processing unit 10 removes all periodic chains of a place where the cycle is shorter than specified proportion of an overall length of a video data in Step S102. That is, the video voice processing unit 10 can eliminate a periodic chain of a short cycle without a chance of expressing a news item, by performing this process. Such a cycle may be generated, when a chairman interviews a guest, for example, or when other short-time cycles appear in newscasting.

[0219]And in Step S103, the video voice processing unit 10 about all the periodic chains which remained in Step S102. When it asks for the time shortest periodic chain and this periodic chain laps with other periodic chains, that periodic chain is removed from a list of periodic chains. The video voice processing unit 10 repeats this processing until it is lost that any periodic chains lap with other periodic chains. A list of periodic chains which remained after this step S103 was completed will include a detected news item list. That is, each cycle of a list of periodic chains obtained at Step 103 expresses one news item, respectively.

[0220]Thus, the video voice processing unit 10 can detect a news item automatically using a chain.

[0221]in addition -- it should mention especially -- the video voice processing unit 10 is being able to act satisfactorily, for example, also when a newscaster changes in the middle of newscasting called between each segment of the main of newscasting, a sport, and business.

[0222]Below, a case where a play in a sportscast is detected is explained.

[0223]Many sports have the feature of having the fixed pattern that a play is constituted, by repeating a series of same processes repeatedly. For example, in the case of baseball, a pitcher throws a ball, and a play is constituted when a batter tries to hit a ball. In a video data, football and Rugby are mentioned as other team sports which have such a play structure, for example.

[0224]When this play structure is broadcast, a video data will express a repetition of a segment group about each portion of a play. That is, when a segment showing a batter continues after a segment with which a video data expresses a pitcher and a ball is hit, a segment showing an outfield player etc. will enter. Therefore, when chain detection by the video voice processing unit 10 is applied to baseball broadcast. In a video data, a segment showing a pitcher will be detected as one chain, one chain with an another segment showing a batter will be occupied, and other chains will hit the outfield and various scene.

[0225]That is, in these sportscasts, play structure serves as a periodic image detectable using a periodic chain detecting method mentioned above. Tennis is mentioned as such other examples. In tennis, a video data constitutes a serve, a volley, a serve, and a cycle like a volley. In this case, since a segment showing each serve is mutually similar pictorially, such a segment can be used for the video voice processing unit 10 in order to detect a play. As a result, in structural analysis by the video voice processing unit 10, play structure of a game is approximately detectable.

[0226]In other sports, especially an individual event, as a play structure, it will carry out until one contestant completes a certain activity, but the whole of each contestant can consider that the same activity is performed approximately. For example, in a ski-jumping game, each contestant performs a jump once, the next contestant continues and the same jump is performed. That is, as for a video data in broadcast of a jump game, it is common for a contestant to start preparation of a jump, and to slide on an in-run, to get down, and to consist of a row of a segment of landing. From this, a video data comprises repeating such a series of segments for every contestant. When chain detection is applied to a video data in such broadcast, a series of chains similar for every stage of a jump will be detected. Therefore, a cycle for every contestant can be extracted using a periodic chain detecting method.

[0227]In the video voice processing unit 10, when chain analysis detects a play in a sportscast automatically, in order to eliminate a chain which is not suitable, it may be

necessary to provide the further restrictions. Although what kind of restrictions are appropriate changes with kinds of sport, an experiential rule of detecting only a thing in which the cycle is sufficiently long among detected periodic chains as a play can be used for the video voice processing unit 10, for example.

[0228]That is, the video voice processing unit 10 performs a series of processings in which an outline is shown in drawing 18, in order to detect a play in a sportscast automatically using a chain.

[0229]First, the video voice processing unit 10 detects a periodic chain in Step S111 using periodic chain detection algorithms mentioned above, as shown in the figure.

[0230]And in Step S112, the video voice processing unit 10 applies sea damaged terms to a list of obtained chains, filters the chained list, and removes a chain which is not essential. Leaving only a periodic chain which is crossed to the great portion of program as sea damaged terms, for example is mentioned. Of course, the video voice processing unit 10 may add constraints peculiar to the target sport.

[0231]Thus, the video voice processing unit 10 can detect a play in a sportscast automatically in chain analysis.

[0232]Below, a case where a topic is detected combining periodic detection and scene detection is explained.

[0233]For example, a video data in many TV programs, such as a drama, a comedy, and variety, is constituted by scene mentioned above. however, a video data comprises a row of some related scenes as a structure of the higher rank -- a topic -- it may have structure. This topic is not necessarily similar with a topic in newscasting which always starts in an introduction segment by a studio chairman. For example, as a visual example, a segment of a logo image or an anchorman's segment may be used instead of an introduction segment, or the theme music always same whenever a new topic starts as an auditory example may be passed.

[0234]It can be judged by combining periodic detection and scene detection whether a video data in a certain program has such a topic structure.

[0235]Therefore, the video voice processing unit 10 performs a series of processings in which an outline is shown in drawing 19, in order to perform topic detection which combined periodic detection which used a chain, and scene detection.

[0236]First, as shown in the figure, in Step S121, the video voice processing unit 10 performs basic similar chain detection, and identifies 1 set of basic similar chained lists.

[0237]Next, in Step S122, the video voice processing unit 10 performs periodic chain detection, and identifies a list of 1 set of periodic chains.

[0238]Then, in Step S123, the video voice processing unit 10 applies an algorithm previously shown in drawing 16 using a basic similar chained list for which it asked in Step S121, and extracts scene structure. As a result, the video voice processing unit 10 can obtain a list of scenes.

[0239]And the video voice processing unit 10 is compared with each scene element which detected a list of periodic chains for which it asked in Step S122 in Step S123 in Step S124. Here, the video voice processing unit 10 removes all periodic chains of a cycle shorter than a scene contained in a list of detected scenes. Although the remaining periodic chains obtained as a result have a scene of some [cycle / each], each of this cycle will be identified as a candidate topic, respectively.

[0240]Thus, the video voice processing unit 10 can perform topic detection by combining periodic detection and scene detection which used a chain.

[0241]The video voice processing unit 10 can also raise accuracy of topic detection by establishing other restrictions and sea damaged terms in Step S124.

[0242]As mentioned above, the video voice processing unit 10 can determine and output various local video structures and/or various global video structures of a video data using detected various chains.

[0243]As explained above, the video voice processing unit 10 shown as an embodiment of the invention can detect a similar chain which comprises two or more mutually similar video segments or sound segments. And the video voice processing unit 10 can extract video structure of a high level by analyzing these similar chains. Especially the video voice processing unit 10 can conduct analysis of local video structure and global video structure by a common framework.

[0244]This video voice processing unit 10 can be processed automatically thoroughly, and a user does not need to know structure of the contents of the video data a priori.

[0245]It is possible also for analyzing video structure sequentially by using successive chain detection, and if the video voice processing unit 10 has still more powerful count ability of a platform enough, it can conduct video structure analysis in real time. Thereby, the video voice processing unit 10 can be used also for video broadcast of the live besides a video data recorded a priori. For example, the video voice processing unit 10 is applicable to a sportscast of the live in play detection in a sportscast.

[0246]The video voice processing unit 10 can give the foundation of new high-level access for video browsing, as a result of detecting video structure. That is, the video voice processing unit 10 enables access to a video data based on the contents by converting the contents of the video data into a video signal using video structure of a high level called not a segment but a topic. For example, by displaying a scene, the user can know a gist of a program quickly and the video voice processing unit 10 can find an interested portion promptly.

[0247]The video voice processing unit 10 enables access of a powerful and new method to newscasting by using a result of topic detection in newscasting further again at a user, such as enabling selection and viewing and listening in a news item unit.

[0248]The video voice processing unit 10 can give the foundation for creating an

abstract of a video data automatically as a result of video structure detection. In order to create a coherent abstract generally, it is required not to combine arbitrary segments contained in a video data, but to decompose into an ingredient with a meaning which can reconstruct a video data, and to combine a suitable segment for origin for it. Video structure detected by the video voice processing unit 10 provides fundamental information for creating such an abstract.

[0249]It is possible to analyze a video data according to the genre in the video voice processing unit 10. For example, the video voice processing unit 10 makes it possible to detect only a game of tennis.

[0250]The video voice processing unit 10 makes it more possible by being included in a video editing system in a broadcasting station to edit a video data based on the contents than this.

[0251]The video voice processing unit 10 can be used for analyzing home video or extracting video structure from home video automatically in an ordinary home further again. The video voice processing unit 10 can be used for performing an abstract of the contents of the video data, and edit based on the contents.

[0252]On the other hand, the video voice processing unit 10 can be used as a tool supplementary to analysis of the contents of the video data according a video chain to a help. Navigation and video structure analysis of the contents of a video data can make easy especially the video voice processing unit 10 by converting a result of chain detection into a video signal.

[0253]Since efficiency of the video voice processing unit 10 in which the algorithm is very simple and calculative is good, it is applicable also to household electronic equipment, such as a set top box, a digital video recorder, a home server.

[0254]Characteristic quantity which this invention is not limited to an embodiment mentioned above, and is used for similarity measurement between segments, for example, the contents of the applicable video data, etc., Of course, except what was mentioned above may be sufficient, in addition it cannot be overemphasized that it can change suitably in the range which does not deviate from the meaning of this invention.

[0255]

[Effect of the Invention]As explained to details above, the signal processing method concerning this invention, It is a signal processing method which detects and analyzes the pattern reflecting the semantic structure of the contents of the supplied signal, The characteristic quantity extraction process of extracting at least one or more characteristic quantity showing the feature from the segment formed from a series of the continuous frame which constitutes a signal, The metrics which measure the similarity between the pairs of a segment are computed for every each of characteristic quantity using characteristic quantity, It has a detection process which detects the similar chain which comprises two or more segments mutually similar

among segments using the similarity measuring process which measures the similarity between the pairs of a segment by these metrics, and characteristic quantity and metrics.

[0256]Therefore, the signal processing method concerning this invention can detect the fundamental structural patterns which a segment similar in a signal constitutes, and can extract the structure of a high level by analyzing how these structural patterns are combined.

[0257]The video voice processing unit concerning this invention is a video voice processing unit which detects and analyzes the pattern of the image reflecting the semantic structure of the contents of the supplied video signal, and/or a sound, The feature amount extracting means which extracts at least one or more characteristic quantity showing the feature from the image formed from a series of the continuous image which constitutes a video signal, and/or an audio frame, and/or a sound segment, The metrics which measure the similarity between the pairs of an image and/or a sound segment are computed for every each of characteristic quantity using characteristic quantity, The similarity measuring means which measures the similarity between the pairs of an image and/or a sound segment by these metrics, It has a detection means to detect the similar chain which comprises two or more images and/or sound segments mutually similar among an image and/or a sound segment, using characteristic quantity and metrics.

[0258]Therefore, the video voice processing unit concerning this invention, It becomes it is possible to determine and output an image similar in a video signal and/or the fundamental structural patterns of a sound segment, and possible to extract the video structure of a high level by analyzing how these structural patterns are combined.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1]It is a figure explaining the composition of the video data applied in this invention, and is a figure explaining the structure of the modeled video data.

[Drawing 2]It is a figure explaining the similar chain which extracts local video structure.

[Drawing 3]It is a figure explaining the similar chain which extracts global video structure.

[Drawing 4]It is a block diagram explaining the composition of the video voice processing unit shown as an embodiment of the invention.

[Drawing 5]In the video voice processing unit, it is a flow chart explaining a series of

processes at the time of detecting and analyzing video structure.

[Drawing 6]It is a figure explaining the amount sampling processing of dynamic features in the video voice processing unit.

[Drawing 7]It is a figure explaining a basic similar chain.

[Drawing 8]It is a figure explaining a link similar chain.

[Drawing 9]It is a figure explaining a periodic chain.

[Drawing 10]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting a basic similar chain using batch clustering technology.

[Drawing 11]It is a figure explaining a dissimilarity nature threshold.

[Drawing 12]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of performing chain filtering of a basic similar chain.

[Drawing 13]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting a basic similar chain using clustering technology one by one.

[Drawing 14]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting a link similar chain.

[Drawing 15]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting a periodic chain.

[Drawing 16]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting a scene using a chain.

[Drawing 17]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting a news item using a chain.

[Drawing 18]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of detecting the play in a sportscast using a chain.

[Drawing 19]In the video voice processing unit, it is a flow chart explaining a series of processes at the time of performing topic detection which combined periodic detection and scene detection using the chain.

[Description of Notations]

10 A video voice processing unit and 11 [A voice feature amount extraction part and 15 / A segment characteristic quantity memory and 16 / A chain primary detecting element and 17 / A characteristic quantity similarity test section and 18 / Chain analyzing parts] A video dividing part and 12 A video segment memory and 13 An image feature quantity extracting part and 14